

Universidade de Lisboa
ISEG Lisbon School of Economics and Management



Uncertainty Quantification with a Gaussian Process Prior

An Example from Macroeconomics

Autor: Ivo Alberto Valente Tavares

Orientador: Prof. Doutor Rui Paulo

Tese especialmente elaborada para obtenção do grau de
Doutor em Matemática Aplicada à Economia e Gestão.

2021

Universidade de Lisboa
ISEG Lisbon School of Economics and Management



Uncertainty Quantification with a Gaussian Process Prior
An Example from Macroeconomics

Autor: Ivo Alberto Valente Tavares

Orientador: Prof. Doutor Rui Miguel Baptista Paulo

Tese especialmente elaborada para obtenção do grau de
Doutor em Matemática Aplicada à Economia e Gestão.

Júri

Presidente: Doutor Nuno João de Oliveira Valério
Instituto Superior de Economia e Gestão da Universidade de Lisboa

Vogais Doutor Hedibert Freitas Lopes
Insper, São Paulo, Brasil

Doutor Gonzalo García-Donato Layron
Department of Economic Analysis and Finance of Universidad de Castilla-La
Mancha, Spain

Doutor Rui Miguel Baptista Paulo
Instituto Superior de Economia e Gestão da Universidade de Lisboa

Doutora Dalia Chakrabarty
Department of Mathematics of College of Engineering, Design and Physical
Sciences at Brunel University

Doutor Giovanni Loiola da Silva
Departamento de Matemática do Instituto Superior Técnico da Universidade
de Lisboa

Abstract

Uncertainty Quantification Using a Discrepancy Term with a Gaussian Process Prior

An Example from Macroeconomics

Ivo Tavares

Abstract

This thesis may be broadly divided into 4 parts.

In the first part, we do a literature review of the state of the art in misspecification in Macroeconomics, and what so far has been the contribution of a relatively new area of research called Uncertainty Quantification to the Macroeconomics subject. These reviews are essential to contextualize the contribution of this thesis in the furthering of research dedicated to correcting non-linear misspecifications, and to account for several other sources of uncertainty, when modelling from an economic perspective.

In the next three parts, we give an example, using the same simple DSGE model from macroeconomic theory, of how researchers may quantify uncertainty in a State-Space Model using a discrepancy term with a Gaussian Process prior.

The second part of the thesis, we used a full Gaussian Process (GP) prior on the discrepancy term. Our experiments showed that despite the heavy computational constraints of our full GP method, we still managed to obtain a very interesting forecasting performance with such a restricted sample size, when compared with similar uncorrected DSGE models, or corrected DSGE models using state of the art methods for time series, such as imposing a VAR on the observation error of the state-space model.

In the third part of our work, we improved on the computational performance of our previous method, using what has been referred in the literature as Hilbert Reduced Rank GP. This method has close links to Functional Analysis, and the Spectral Theorem for Normal Operators, and Partial Differential Equations. It indeed improved the computational processing time, albeit just slightly, and was accompanied with a similarly slight decrease in the forecasting performance.

The fourth part of our work delved into how our method would account for model uncertainty just prior, and during, the great financial crisis of 2007-2009. Our technique allowed us to capture the crisis, albeit at a reduced applicability possibly due to computational constraints. This latter part also was used to deepen the understanding of our model uncertainty quantification technique with a GP. Identifiability issues were also studied. One of our overall conclusions was that more research is needed until this uncertainty quantification technique may be used in as part of the toolbox of central bankers and researchers for forecasting economic fluctuations, specially regarding the computational performance of either method.

Keywords: Uncertainty Quantification, Machine-learning, Misspecification, Non-Linearities, Structural Uncertainty, Model Discrepancy, Gaussian Processes, Hilbert Methods, Financial Crisis, State-Space Models.

Dedication

To God

Acknowledgements

I acknowledge the help of my thesis advisor Professor Rui Paulo who helped me improve my work, and with whom I had valuable conversations.

I would also like to acknowledge the help of Professor João Janela who helped me work with CEMAPRE's servers, which were a valuable resource, without which I could not have reached many of my results.

Lastly, I also acknowledge the help Professor Nuno Sobreira gave me at the beginning of my work in discovering some bibliography.

Abbreviations

AR - Autoregression
AS - Active Space
ASD - Agnostic Structural Disturbances
ASGP - Active Space Gaussian Process
BMA - Bayesian Model Averaging
BVAR - Bayesian Vector Autoregression
BCS - Blue Chip Survey
DFM - Dynamic Factor Model
DGP - Data Generating Process
DP - Dynamic Programming
DSGE - Dynamic Stochastic General Equilibrium
GP - Gaussian Process
FEVD - Forecast Error Variance Decomposition
FRED - Federal Reserve Economic Data
HRRGP - Hilbert Reduced Rank GP
IRF - Impulse Response Function
IS - Importance Sampling
LRE - Long-run expectations
MCMC - Markov Chain Monte Carlo
MH - Metropolis-Hastings
MOGPR - Multi-Output Gaussian Process Regression
PD - Positive Definite
PG - Particle Gibbs
PGAS - Particle Gibbs Ancestor Sampling
RBC - Real Business Cycle
RMSE - Root Mean Square Error
RWMH - Random Walk Metropolis-Hastings
SIS - Sequential Importance Sampling
SMC - Sequential Monte Carlo
SPD - *Symmetric* Positive Definite
SSM - State-Space Model

SVAR - Structural Vector Autoregression

SW - Smets & Wouters

UQ - Uncertainty Quantification

VAR - Vector Autoregression

VFI - Value Function Iteration

Contents

1	Introduction	10
2	Towards Uncertainty Quantification in Macroeconomics	16
2.1	Definition of UQ	16
2.2	Inverse Uncertainty Quantification: A Bayesian Perspective . .	19
2.3	Uncertainty Quantification in Macroeconomics	21
2.3.1	A Global Sensitivity Analysis for Parametric Uncertainty	21
2.3.2	Using Gaussian Processes for a Global Surrogate Model and Parameter Uncertainty Quantification	24
2.4	Uncertainty Quantification in Computer Models: Using GPs .	27
2.4.1	A Methodology for Recognizing Several Sources of Un- certainty	27
2.4.2	Identifiability Issues and Possible Solutions	31
2.5	Contributions of Our Work	35
3	Dealing with Misspecification in Macroeconomics	38
3.1	Dynamic Stochastic General Equilibrium Models (DSGE) . . .	38
3.2	Hybrid Models	40
3.2.1	Additive Hybrid Models	40
3.2.2	Hierarchical Hybrid Models	42
3.3	Bayesian Model Averaging	45
3.4	Recent Advances	46
3.4.1	Identification of misspecification in Data-Rich Envi- ronments	46
3.4.2	Identification of misspecification by introducing uncor- related exogenous processes	47
3.4.3	Using A Composite Likelihood approach	49
3.4.4	Agnostic Structural Disturbances	50
4	UQ of Model Discrepancy with a Full Gaussian Process Prior	53
4.1	The Landscape	53

4.2	Quantifying Structural Uncertainty in a DSGE Model	54
4.2.1	A Bias Term with a Gaussian Process Prior	54
4.2.2	Prediction in a DSGE Model with a GP Bias Term	56
4.3	Learning the Model	58
4.3.1	Sampling θ from its conditional posterior distribution	58
4.3.2	Sampling from the State-Space	60
4.3.3	The DSGE Model	62
4.3.4	Covariance Function, Priors and Proposal	68
4.4	Results	72
4.4.1	A Preliminary Simulation	72
4.4.2	Comparing Performances	87
4.5	Conclusions	88
5	UQ using a Hilbert Reduced Rank Approximation to a GP	90
5.1	Introduction	90
5.1.1	Hilbert Space Methods for a Reduced-Rank Multidimensional GP	91
5.1.2	Hilbert Reduced-Rank GP in a State-Space Model with a Discrepancy Term	93
5.1.3	Prediction in a SSM with a HRRGP Prior on a Discrepancy Term	94
5.1.4	Sampling from the Parameters	95
5.1.5	Sampling from the State-Space	97
5.2	Covariance Function, Priors and Proposal	98
5.2.1	Covariance Function	98
5.2.2	Priors	99
5.2.3	Proposal Distribution	100
5.3	Results	101
5.3.1	Preliminary Simulation	101
5.3.2	Comparing Performances	108
5.4	Conclusions	108
6	UQ during the Great Crisis/Recession	110
6.1	The Great Crisis: A Summary	110
6.2	DSGE Models during the Great Crisis	111
6.3	Can our UQ framework capture the Crisis?	113
6.3.1	Data Plots	114
6.3.2	Traces	115
6.3.3	Posterior Histograms and Density Priors	119
6.3.4	Forecasts	123
6.4	Some Tighter Priors, and a Smoother Covariance Function	125

6.4.1	Results For Matèrn Smooth Covariance Function . . .	126
6.4.2	Results For Squared Exponential Covariance Function .	132
6.5	Conclusions	137
7	Conclusions for the Present and the Future	139
A	A Brief Introduction to Gaussian Processes	142
A.1	Univariate Gaussian Processes	142
A.1.1	Function-Space View	142
A.1.2	Weight-Space View	143
A.1.3	Covariance Functions	144
A.1.4	Multi-Output Gaussian Process Regression	146
B	Bayesian Statistical Methods	148
B.1	Distributions for HRRGP	148
B.1.1	Matrix-Normal Distribution	148
B.1.2	Inverse-Wishart Distribution	149
B.1.3	The Matrix-Normal Inverse-Wishart Conjugate Prior .	149
B.2	Introducing Particle Filtering Algorithms	151
B.2.1	Monte Carlo Method	151
B.2.2	Importance Sampling	152
B.2.3	Sequential Importance Sampling	153
B.2.4	Particle Filtering	155
B.3	Using a PGAS for y_M	159
B.3.1	Trace Plots	161
B.4	The PGAS Markov Kernel for a HRRGP	162
C	Hilbert Space Methods for Reduced-Rank Gaussian Pro-	
	cesses	164
C.1	Kernels as Integral Operators	164
C.1.1	Stationary Kernels	165
C.1.2	Eigenfunction Analysis of Kernels	165
D	Derivations for Langevin type of Metropolis-Hastings Step	167
E	Numerical Considerations	176
E.1	Alternative to Inverting a Matrix	176
E.2	Guaranteeing Positive Definiteness and Symmetry	176
	Bibliography	178

Chapter 1

Introduction

An uncertainty quantification(UQ) analysis of a model has as a main objective the quantification of, and possible accounting for, different sources of uncertainty that arise in the use and assessment of the underlying scientific model. As the scientific theory progresses, better models are created, but usually there is also an increase in complexity. However, once a very complex model is established, some modelling aspect may still lack, leading to model misspecification. In some instances, parameters estimation may overfit due to model misspecification, and scientific interpretation may be lost. In Macroeconomic literature, the issue of misspecification, or structural uncertainty, is perennial; with this work, we hope to contribute to the furthering of the area. Some good introductions to the topic of UQ can be found in references Sullivan (2015) and Soize (2018).

Two views: Empirical and Theoretical

Broadly speaking, in Macroeconomics, one observes two, at a time opposing or complementary, views on how to analyse the fundamental variables of interest. On the one hand, we have a more econometric style of analysis, where models such as Structural Vector Auto-Regressive(S-VAR) or Bayesian VARs are used, models that can be considered statistical in nature, originating from time series. On the other hand, especially since Lucas (1976)'s critique to econometric empirical models for not being invariant to policy changes (leading to great instability in the estimation of such models across time), other types of models for macroeconomic analysis were developed, and they were called Dynamic Stochastic General Equilibrium (DSGE) models. The objective was to capture those eluding structural parameters, and still retain some explanatory power of the macroeconomic data. However, due to severe simplifications, and a stringent adherence to its theoretical assumptions, many

DSGE models do not hold their own against econometric methods in terms of predictive performance.

The Great Moderation, a.k.a the honeymoon period for DSGE

DSGE modelling is the product of the effort of macroeconomists to endogenize¹ the expectations of agents (households, firms, etc.) in the economy while optimizing their behaviour. Whereas previous macroeconomic models approach to data was based on ad-hoc behavioural rules, DSGE models rely on micro-foundations. These were the answers found by macroeconomists to the Lucas critique of the 1970s, in which ad-hoc exogenous behavioural rules were found to be inconsistent with behavioural choices from rational agents² in response to policy shifts. The appeal of these highly formal(albeit some may say artificial) models, together with the *Great Moderation*³ of economic performance, enthralled researchers in the academy and central banks alike. In Lucas (2003), authored by Robert Lucas, one can read:

My thesis in this lecture is that Macroeconomics in this original sense has succeeded: Its central problem of depression prevention has been solved, for all practical purposes, and has in fact been solved for many decades.

During this period, Central Banks incorporated them in their broad range toolkit, which comprehended also other models such as SVARs, as an aide to policy making, some even developing their own specific DSGE models such as SIGMA of the US Federal Reserve system, or the European Central Bank's NAWM (see Tovar (2008)).

The Great Recession

The advent of the *Great Recession* of 2007-2009 contributed to putting a dent on the confidence researchers had on these models. Some of the leading macroeconomic researchers wrote, through out the years, scathing criticisms

¹The term *Exogenous* describes variables whose behaviour are not determined by other variables from the scientific model. We could consider them as determined outside from the scientific model. Its antonym is *Endogenous*.

²Rational agents are intervening model actors whose behaviour follows an axiomatic set on its utility, which supposedly characterises rationality. Lately, several findings in behavioural economics have raised serious doubts on the validity and applicability of this axiomatic set.

³Term is used to describe a period starting in mid 80s, characterised by a marked decrease in the volatility of business cycle (GDP, industrial production, etc.) fluctuations in the developed countries. This period ended with the financial crisis of 2007-2009.

to a naive use of DSGE models, see for example Romer (2016), a corrosive paper by Paul Romer, or Blanchard (2016), an ominously entitled paper "Do DSGE models have a future?". In our opinion, a balanced representative description of the critics opinions is given by Olivier Blanchard's bleak perspective in Blanchard (2018), where he published his long matured thought about DSGE models, by writing

I am not optimistic that DSGEs will be good policy models unless they become much looser about constraints from theory. I am willing to see them used for forecasting, but I am again sceptical that they will win that game.

In fact, in the same article, the author calls for five different classes of models to be distinguished, giving up on the scientific cointegration objective where the same model would be used for several different goals.

Foundational models make a deep theoretical point, likely of relevance to nearly any macro model, but not pretending to capture reality closely. An example of such models may be the overlapping generation model of Diamond (1965).

DSGE models explore the macro implications of distortions. Supposedly built around a largely agreed upon common core, they should try to resemble reality, but not at the cost of adding some ad-hoc repairs with higher order computational burden.

Policy models help policy, and study the dynamic effects of specific shocks, allowing for the exploration of alternative policies. A reflection of the actual dynamics of the economy is essential for this type of models. An example would be the FRB/US model used at the Fed.

Toy models give a succinct explanation to a question, representing the essence of the answer from a more complicated model or from a class of models. Adherence to strict theory is not a relevant criteria for these models. Some examples are the IS-LM model, the Mundell–Fleming model, the RBC and the New-Keynesian models.

Forecasting models give the best forecasts, and it is the only criteria on which to judge these models. Most time-series econometric models fall in this category. Again, adherence to strict theory is not a relevant criteria for these models. If theory helps, it should be used. Otherwise, it should not...

Whether theory helps forecasting performance, Blanchard says it is still too early to tell.

Christiano, Eichenbaum, and Trabandt (2018) presents a vitriolic⁴ defence of DSGE models; they describe the main areas of research attempting at creating more robust DSGE models in the aftermath of The Financial Crisis. The main ones were the addition of financial frictions, the zero lower bound condition on interest rates, policy analysis with non-linearities, and incorporation of heterogeneous agents (agents with different borrowing constraints). The furthering of the scientific theory to improve modelling concerns may address some sources of model misspecification, but is certainly not the only avenue of possible research, as was acknowledged in Negro and Schorfheide (2009), which uses the DSGE-VAR approach to account for model misspecification. The interested reader may consult, in the present work, Chapter 3, subsection 3.2 entitled "Hybrid Models".

More information on how certain researchers of Central Banks currently perceive the usefulness of DSGE models may be consulted on the reference Gürkaynak and Tille (2017). In it, J.C. Williams, from the Federal Reserve Bank of San Francisco, defends their use as part of a broader toolkit ranging from empirical to theoretical/formal models, possibly hinting at the use of a methodology akin to Bayesian Model Averaging. Del Negro and Giannoni, from the Federal Reserve Bank of New York, acknowledge the limitations in forecasting performance of DSGE models, but also highlight one of their greatest strengths - at least theoretically - the experiments on policy counterfactuals. Since the behaviour of agents has already been endogenized (completely incorporated in the model), the solution of the model is invariant to policy changes. Hence, a central bank researcher may study the path of the economy under different policy decisions. They finish by concluding that greater contact between model developers at Central Banks and academic researchers may improve the development of more robust DSGE, and state⁵

One way to guard against model misspecification is to look across types of DSGE models, possibly also including reduced-form models, with weights on the models in these pools of models possibly varying over time depending on the question at hand

⁴In earlier versions of this reference: "people who don't like DSGE models are dilettantes". The authors later removed this sentence, and toned down a bit. In Muellbauer (2018), one may still find a reference to this fact.

⁵For a resemblance between this quote and the Bayesian Model Averaging (BMA) method, see the corresponding section in this work.

However, at any point in time, regardless of how complex economic models are, there can still be some serious misspecification, given reality’s complexity. Hence, a complementary approach to improving the economic theory, is to directly tackle the problem of misspecification.

Our framework

Tackling the problem of misspecification resulted in a way of integrating both views, empirical/econometric and theoretical, with the creation of the so-called Hybrid Models. Take a DSGE model in its state-space reduced form representation, and add an observation error term, which follows a VAR model dynamic, to its observation equation, and the end result is what is called an *additive* hybrid model. A short introduction to hybrid modelling can be found in Schorfheide (2013). It is on these additive type of models that our framework bears its resemblance to. Instead of adding a simple VAR error term, we will add a much more complex term, having a Gaussian Process prior, which we call discrepancy term, in hope to absorb that behaviour which the original scientific model cannot capture. Our idea of adding a discrepancy term with a Gaussian Prior comes from Kennedy and O’Hagan (2001). Although in that paper the method was applied to engineering models — such as the radiation gaussian plume model — and in a regression setting, we intend to show how, in an area such as Macroeconomics, where no experimentation strictly-speaking is possible, such a method may be applied, in a state-space setting. This method has been applied extensively in the area of computer experiments, i.e. the study of computational models used to replicate real, physical systems, which encompass computational physics, computational chemistry, computational biology, and other similar subjects. We will consider the macroeconomic model as the ‘emulator’, in our work a source of uncertainty by being a simplification of the true physical system(‘computer model’). Methods using Gaussian Priors have lately abounded in the machine learning literature, and in fact the main bibliographical references for this work belong to that subject, see for example M. A. Alvarez, Rosasco, and Lawrence (2012), Frigola-Alcalde et al. (2013), Lindsten, Jordan, and Schön (2014), H. Liu, J. Cai, and Ong (2018), Rasmussen and Williams (2006), Solin and Särkkä (2014), and Svensson et al. (2016) .

In the DSGE literature, one can find some examples where we can observe how an additive hybrid model has improved forecasting performance, and obtain structural parameter estimates more in line with previous scientific studies, one of the most eminent reference being Ireland (2004). Hence, inspired by all these references, our method is to add a model discrepancy,

or bias term, with a Gaussian Process prior to correct the misspecification of the DSGE, whilst hoping to obtain a better interpretability of the structural parameters. The bias term, with the Gaussian Process prior, is a non-parametric term which may be sufficiently flexible to capture that source of uncertainty. Hence the relation to UQ, a relatively new area of science where we try to quantify and account for the sources of uncertainty in a model. However, there is literature documenting cases where, even with somewhat simpler models, a non-parametric discrepancy term such as this may lack the ability to overcome the gap between the computer model and the physical model, such as in Brynjarsdóttir and O’Hagan (2014).

The Plan

The remaining part of this work is organized as follows. In Chapter 2, we describe in greater detail what is UQ, its connections to inverse problems, and what has been published in Macroeconomics related to UQ. In Chapter 3, we survey on how macroeconomic researchers have proceeded in their objective of correcting for misspecification, a known source of model bias, and in what sense the method we propose differs, and may allow for greater flexibility by not assuming a specific parametric model, or even linearity in the scientific model. In Chapter 4, we develop our framework, giving the model discrepancy term a full multi-output GP prior, and analyse its performance for the same data set as in Ireland (2004). In Chapter 5, we use Hilbert Space techniques to find a faster approximation to the full GP of Chapter 4. We also use the same data set for assessing its predictive performance. Then, in Chapter 6, we apply our framework to the period of the financial crisis, and analyse its predictive performance. And finally, in Chapter 7, we give an overall conclusion of our work, consolidating all the previous chapters.

Chapter 2

Towards Uncertainty Quantification in Macroeconomics

2.1 Definition of UQ

Uncertainty Quantification is a relatively new scientific subject being developed with the objective of characterizing and reducing the sources of uncertainty¹ in modelling real world and computational applications. An example of such an application may be the study of an optimal airfoil shape for an aircraft, where certain physical parameters may be uncertain since no physical object is exactly equal to another, nor is every try in an experiment. Furthermore, if these experiments take too much time, then the engineer may often resort to constructing approximation/surrogate models. Another example of application of UQ is computer experiments where the physical system — which may be the entire earth as in climatology — due to its complexity must be simulated using a computer. In this situation, some simulators may impose a very high computational burden, limiting the amount of 'observed'(simulated) data.

In both the climatological and engineering examples, the presence of uncertainty due to our limited information may call for the use of UQ techniques for a proper statistical analysis. The modelling of an experiment (real or simulated) may introduce uncertainty from several possible sources such as²

¹We can interpret uncertainty here as a kind of knowledgeable ignorance, i.e. we know that we don't know a certain information

²The following does not intend to be an exhaustive list of the sources.

Parameter uncertainty The researcher ignores the exact true values that should be inputted into the model being used.

Parametric variability We may know the exact general theoretical values to be used, but due to the specificities of our experiment, these may have changed or be unknown. For example, we will not measure a million coins by hand, instead it is much easier to just use the factory dimensions which are given. The factory settings, which are known, may not be the same as the settings our object has which are unknown to us, due to physical constraints.

Structural uncertainty Also known as model discrepancy, bias or inadequacy, it originates from a lack of knowledge of the true system, where an approximate model is used. Suppose a linear model is used to study a non-linear reality. In a sense, it is similar — if not equal — to the concept of model misspecification³.

Algorithmic uncertainty Also known as numerical uncertainty or discrete uncertainty, may come from limitations in our numerical solving methods, which introduce errors, and only give approximate solutions.

Interpolation Uncertainty Originating from a shortage of available data, pushing the researcher to interpolate or extrapolate in order to predict the model responses. For example, imagine a theoretical method allows to find an approximation to a point where we have no data, within an error bound. This is an example of interpolation uncertainty. However, if some of the theoretical steps are hindered by computational power limitations, we may have to employ numerical solving techniques which may add a non-negligible source of algorithmic uncertainty.

Experimental Uncertainty Can also be interpreted as measurement error, and an ever present source of uncertainty in experiments.

Although there may be other sources of uncertainty, those presented above are usually the main ones. Let us now consider the following equation

$$y^{\text{obs}} = y^{\text{eco}}(x, \theta^*) + \delta(x) + \epsilon$$

³By model specification our understanding is the construction/decision for a functional form to be taken as the model. When this functional form is inappropriate, it is a source of model bias. Farther in the text, we will also ponder its connections with other types of bias, namely omitted-variable bias.

From the above equation, we can already exemplify some sources of uncertainty. In a sense, reality is the true data generating process which is only observed in a very noisy or uncertain way. On the left side of the equation we have the term y^{obs} which represents our observations or experimental responses. The most easily recognized source of uncertainty is experimental uncertainty (measurement errors), and these are usually modelled by the additive term ϵ . However, there may be other sources. With θ^* , we are representing some parameter values which should be used, but are usually unknown, and thus this is a source of uncertainty related to parametric uncertainty and parametric variability. A third source of uncertainty, and which we will delve into, in this work, is model discrepancy or structural uncertainty. This may come from a misspecified or approximate economic model being used, and in the equation it is represented by y^{eco} . Our attempt at capturing that misspecification is by adding a model discrepancy term, represented by δ above. Whether the error term ϵ is able to capture all the possible sources of experimental uncertainty or not, i.e. well-specified, may determine if what it is quantified by the discrepancy term is strictly due to the structural uncertainty/model bias or also to measurement error. In this sense, it does not diverge much from what can happen when using solely the economic model, if measurement errors are not well captured by the ϵ term.

Uncertainty Quantification problems may be divided into two different main branches/perspectives: Forward Propagation of Uncertainty and Inverse Uncertainty Quantification.

Forward Propagation of Uncertainty This is the problem of quantifying how the uncertainty of inputs will propagate forward to uncertainty of outputs. It usually focuses on the assessment of parametric variability as a source of uncertainty.

Inverse Uncertainty Quantification Here, we have the inverse problem, i.e. given output uncertainties, how does one quantify input uncertainties. Whereas the previous problem focused on parameter variability, this type of inverse problems focus on model and parametric uncertainty as some of the main sources of uncertainty.

The subject of Uncertainty Quantification has been deeply interconnected, since its inception, to the analysis of computer models or computer experi-

ments, where surrogate models or emulators would be used instead of very complex computer models, which would introducing new sources of uncertainty. It also has profound connections to geostatistics, where the use of statistical emulators with GP priors, known as *kriging*, has been widely disseminated as a way to account for interpolation uncertainty among other sources. The reader interested in delving deeper into the connections between UQ, computer models, and kriging may find references Rasmussen and Williams (2006) and Soize (2018) useful.

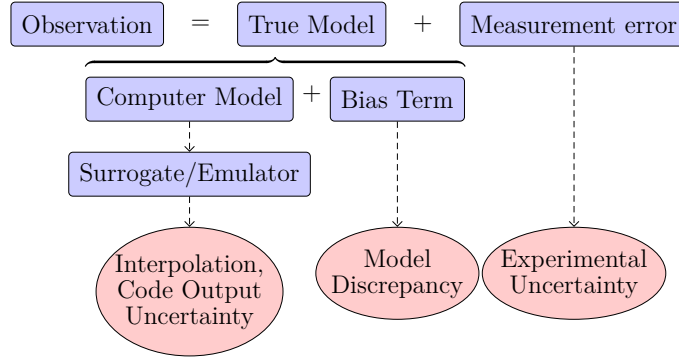


Figure 2.1: UQ in Computer Modelling

2.2 Inverse Uncertainty Quantification: A Bayesian Perspective

An example of solving an Inverse Problem is to find u , a variable or parameter input in a scientific model, given the data/observation y :

$$y = \mathcal{M}(u)$$

with $u \in X$, $y \in Y$, and X, Y Banach spaces. Usually these inverse problems are ill-posed, i.e. may have no solution, may not be unique and may depend sensitively on y (instability).

An approach to tackle this problem of ill-posedness is to reformulate the inverse problem as a least-squares optimization problem^{4,5}

$$\arg \min_{u \in X} \frac{1}{2} \|y - \mathcal{M}(u)\|_Y^2$$

⁴The constant $\frac{1}{2}$ serves the purpose of making easier to establish the connection to a Gaussian distribution, which will be seen in a short moment.

⁵We are using the following notation: $\|a\|_M = a^\top M^{-1}a$

However, this optimization problem by itself may also be ill-posed. A common approach to alleviate this issue is to consider instead a *regularized* optimization problem of the form:

$$\arg \min_{u \in E} \frac{1}{2} \|y - \mathcal{M}(u)\|_Y^2 + \frac{1}{2} \|u - u_0\|_E^2$$

where $E \subset X$ is also Banach, and $u_0 \in E$.

In the classical approach, the choice of norms $\|\cdot\|_E, \|\cdot\|_Y$ and u_0 may be somewhat arbitrary without further specification.

For a Bayesian approach, it is useful to consider the data y as observed with some noise:

$$y = \mathcal{M}(u) + \eta$$

Let us assume that $\eta \sim N(0, B)$ and u has as prior $N(u_0, \Sigma_0)$. One can deduce the posterior distribution to be⁶

$$\begin{aligned} p(u | y) &\propto \exp \left(-\frac{1}{2} \|B^{-1/2}(y - \mathcal{M}(u))\|_2^2 - \frac{1}{2} \|\Sigma_0^{-1/2}(u - u_0)\|_2^2 \right) \\ &= \exp \left(-\frac{1}{2} \|(y - \mathcal{M}(u))\|_B^2 - \frac{1}{2} \|(u - u_0)\|_{\Sigma_0}^2 \right) \end{aligned}$$

Minimizing the original regularized problem would be equivalent to maximize the posterior distribution of our observation equation with Gaussian errors, which will result in the *maximum a posteriori*(MAP) estimator. Another situation where a Bayesian perspective is helpful is when we have an *underdetermined* system. Let us assume that $X = \mathbb{R}^N$ and $Y = \mathbb{R}^J$, and our inverse problem is such that $N > J$ and so the system is *underdetermined*, i.e. the number of unknowns is greater than the number of responses/equations, and by adopting a bayesian framework, the researcher can fill-in the missing information using prior distributions. An example of such a situation is the *Elliptic Inverse Problem*. For more information on the advantages of using a Bayesian framework, which is outside the scope of our work, we direct the interested reader to consult Stuart (2010), Sullivan (2015) and Dashti and Stuart (2015) from where this subsection was inspired from.

Also, since in the Bayesian perspective we have transformed our inverse problem into one of finding a posterior distribution, we can now obtain probabilities of our predictions lying in certain specified regimes, resulting from a characterization of $p(\text{parameters} | \text{data})$ and of $p(\text{future data} | \text{data})$.

⁶ $\|a\|_2$ is the Euclidean norm, such that $\|a\|_2^2 = a^\top a$.

2.3 Uncertainty Quantification in Macroeconomics

Currently, and to the best of our knowledge, the uncertainty quantification carried out in macroeconomics has been mostly centered on a somewhat local sensitivity analysis⁷, namely in the input parameters. Sensitivity analysis may be more broadly understood as understanding how $\mathcal{M}(u_1, \dots, u_n)$ will behave with individual, or *combined*, perturbations in the u_i , i.e. how the uncertainty in the overall model can be assigned to different sources of uncertainty in its inputs. Only recently has there been a greater integration of techniques of the formal subject of UQ and macroeconomics. In the following subsections, some examples are given.

2.3.1 A Global Sensitivity Analysis for Parametric Uncertainty

Harenberg et al. (2019), a variance-decomposition based on Sobol’s indices for quantities of interest is used to order the parameters according to their impact. Some sensitivity experiments take the approach of individually changing the parameter values. These procedures are highly subjective regarding the choice of which parameters to change, i.e. valid only locally, and do not account for possible interactions between different parameters and their relationship in a precise way.

The reference Harenberg et al. (2019) proposes a global sensitivity analysis which overcomes all the limitations stated above. In it, the authors call a computational model a mapping:

$$\theta = (\theta_1, \dots, \theta_M) \in D_\theta \subset \mathbb{R}^M \rightarrow y = \mathcal{M}(\theta) \in \mathbb{R}^Q$$

where y is the *quantity of interest* (QoI), such as moments, or a ratio of some kind.

The *uncertainty propagation* objective is to characterize some stochastic properties of the computational model, e.g. moments or the density associated to y , by regarding the parameters as inputs, and study their contribution to the randomness of our QoI, and in this sense it is a method representing a forward perspective of an uncertainty quantification analysis. The method proposed below is global, since it allocates to each *input parameter* their

⁷Local sensitivity analysis is the study of how the uncertainty of the response/model function can be assigned to small perturbations, in a certain neighbourhood(local). Some few exceptions to this local analysis can be found, such as Ratto (2008), but even here, only individual contributions are considered.

respective contribution to σ_y^2 , the QoI variance. Usually for the stochastic characterization of those sought after properties, Monte Carlo methods were used, which for computationally burdensome models, makes such a global analysis prohibitive. Hence the authors chose to represent the model as a Polynomial Chaos Expansion (PCE)

$$Y = \sum_{j=0}^{\infty} b_j Z_j$$

in which the terms $\{Z_j\}_{j \in \mathbb{N}_0}$ are multivariate orthonormal polynomials - which form the basis of a Hilbert Space - with respect to the *joint distribution function* for θ , i.e. the weights b_j are determined by joint density of θ , f_θ .

The spectral expansion allows us to write

$$Y = \sum_{\alpha \in \mathcal{A}} b_\alpha \Psi_\alpha(\theta)$$

with $\alpha = (\alpha_1, \dots, \alpha_M)$ being a multi-index that identifies the degree of the polynomial for each input parameter θ_i , Ψ_α being a multivariate orthogonal polynomial built by tensor product of the univariate polynomials of degree α_i and with \mathcal{A} being the truncation set. The weights b_α are computed by a least-square minimization based approach by finding the coefficients such that minimize

$$E \left[\left(\mathcal{M}(\theta) - \sum_{\alpha \in \mathcal{A}} b_\alpha \Psi_\alpha(\theta) \right)^2 \right]$$

where the quantity above is integrated over the parameter space. However, to handle computationally with this expression is too complex, and thus we usually approximate it by an empirical mean sampled at a set $\mathcal{X}_{ED} = \{\theta^i, i, \dots, N\}$ called experimental design. The estimates are given by the usual formula

$$\hat{b} = (A^\top A)^{-1} A^\top \mathcal{Y}$$

where $\mathcal{Y} = (\mathcal{M}(\theta^1), \dots, \mathcal{M}(\theta^N))$ and $A = [A_{ij} = \Psi_j(\theta^i), i = 1, \dots, N; j = 1, \dots, \#\mathcal{A}]$ is called the information matrix.

Now, with our approximation⁸ to the model determined by $\hat{Y} = \sum_{\alpha \in \mathcal{A}} \hat{b}_\alpha \Psi_\alpha(\theta)$, it can be shown that $\hat{\mu}_y = \hat{b}_0$ and $\hat{\sigma}_y^2 = \sum_{\alpha \in \mathcal{A}, \alpha \neq 0} \hat{b}_\alpha^2$.

The *Sobol decomposition* states that for any square integrable function \mathcal{M} with $f_\theta = \prod_i f_{\theta_i}(\theta_i)$, we may rewrite

⁸This approximation can also be seen as a *surrogate* model, i.e. an easier to evaluate approximation of the original model.

$$\begin{aligned}
\mathcal{M}(\theta) &= \mathcal{M}_0 + \sum_{i=1}^M \mathcal{M}_i(\theta_i) + \sum_{i < j} \mathcal{M}_{ij}(\theta_i, \theta_j) + \cdots + \mathcal{M}_{12 \dots M}(\theta) \\
&= \mathcal{M}_0 + \sum_{u \subset \{1, \dots, M\}, u \neq \emptyset} \mathcal{M}_u(\theta_u)
\end{aligned}$$

where \mathcal{M}_0 is a constant, and in the last equality we use a set notation, and θ_u is a subvector of θ determined by u . By assuming orthogonality of $\mathcal{M}_u(\theta_u)$ and $\mathcal{M}_v(\theta_v)$, i.e.

$$E[\mathcal{M}_u(\theta_u)\mathcal{M}_v(\theta_v)] = 0, \forall u, v \subset \{1, \dots, M\}, u \neq v$$

we obtain $E[\mathcal{M}_v(\theta_v)] = 0 \forall v \subset \{1, \dots, M\}$.

Using the above properties, together with the existence and uniqueness of the Sobol decomposition, it can be shown that

$$D := \text{Var}[\mathcal{M}(\theta)] = \sum_u D_u$$

with $D_u = \text{Var}[\mathcal{M}_u(\theta_u)] = E[\mathcal{M}_u^2(\theta_u)]$

The Sobol indices S_u are defined as the ratio of the variance of polynomials we are interested in, by the total variance D , i.e. $S_u = \frac{D_u}{D}$. The first-order indices

$$S_i = \frac{\mathcal{M}_i(\theta_i)}{D}$$

and the second-order indices

$$S_{ij} = \frac{\mathcal{M}_{ij}(\theta_{ij})}{D}$$

and so on. The *total* Sobol index

$$S_i^T = \sum_{u: i \in u} S_u$$

quantifies the total effect of parameter θ_i , including interactions with the remaining input parameters. In practice, if we have $S_i^T < 1\%$, the contribution of the parameter θ_i can be considered inconsequential. One interesting consequence is that if a parameter is negligible, then the QoI is not affected by this input parameter, and so even if we have more data on QoI, it will not determine the value of the input parameter, i.e. it is not identifiable. Therefore, a relevant Sobol index is a necessary condition for parameter identifiability, albeit not a sufficient one, since even with a very relevant index one may have a non-linear relation giving rise to several local maxima in the likelihood function. One advantage of PCE is allowing the analytical derivation of these Sobol indices, using only the \hat{b} coefficients. Therefore, not only do we have a surrogate model, i.e. an easier to solve model, but we can also derive analytically the Sobol indices.

2.3.2 Using Gaussian Processes for a Global Surrogate Model and Parameter Uncertainty Quantification

With the 2008 Great Recession, economic researchers turned their focus to many important economic events which could only be studied going beyond local linear dynamics around steady-states or simple representative agents models. Hence, economic models started to grow considerably in complexity - adding heterogeneity, using global solution methods, etc. - to the point where current standard toolkit has been severely challenged by them.

In Scheidegger and Bilonis (2019), the authors propose a global solution method⁹, using Gaussian Processes(GP)¹⁰, together with a parallelization of the Active Space (AS) methodology, to create a surrogate model capable of handling several hundreds of dimensions.

The economic formulation, which will determine the final model, is done using the Dynamic Programming (DP) principle, even though the authors assure us that it could also be applied to a Lagrange Multipliers formulation.

The typical DP formulation of an economic optimization problem consists in maximizing the *value function*

$$V(x_0) = E_0 \left[\sum_{t=0}^{\infty} \beta^t r(x_t, \xi_t) \right]$$

by finding a sequence of controls $\{\xi_t\}_{t=0}^{\infty}$, with $x_{t+1} \sim F(x_t, \xi_t)$, where F is a given transition function, $x_0 \in \mathcal{X}$, $\xi_t \in \Pi(x_t)$ with $\Pi(x_t)$ being the space possible choices of ξ_t , and $r(\cdot)$ being the return function.

Dynamic programming also allows us to find a time-invariant *policy function* p such that $\forall t \in \mathbb{N}$, we have $\xi_t = p(x_t) \in \Pi(x_t)$, while solving the above maximization problem.

The main principle of DP, which is called *principle of optimality*, states that we can find a solution to our optimization problem above by solving the *Bellman Equation*:

$$V(x) = \max_{\xi} \{r(x, \xi) + \beta E[V(x_{t+1})]\}$$

It can be shown that, under certain conditions, the *Bellman Operator*

$$(TV)(x) := V(x)$$

⁹A Global Solution Method is understood as a method which computes a solution by using equilibrium conditions at several points in the state-space, instead of only at a steady-state.

¹⁰See Appendix on GP for a brief introduction.

is a contraction, and hence successive iterations will result in convergence to a fixed-point. This recursive solving is called Value Function Iteration(VFI).

Let us assume $f : \mathbb{R}^D \rightarrow \mathbb{R}$ takes an input x , and responds with an output $f(x)$. We observe only noisy responses, as $t^i = f(x^i) + \epsilon_i$. Instead of using a computationally expensive response surface f , the objective is to use a surrogate model by imposing a GP prior on f and doing a GP regression. For the prior of f , they assign a GP with a square exponential covariance function. This covariance was chosen since it determines the GP to be infinitely differentiable in the mean squared sense, and the economic model when solved by DP require derivatives to be inputted into an optimization routine for the evaluation of the Bellman operator. For the creation of the surrogate of $f(x_t)$, we use some training inputs $X = \{x^i\}$, and output targets(or QoI) $t = \{t^i\}$. Now, we need to assume how the QoI are produced from the outputs. The authors simplify the problem by assuming that each observation is independent from each other, and $t^{(i)} \mid f(x^{(i)}), s_n \sim N(t^{(i)} \mid f(x^{(i)}), s_n^2)$, with s_n^2 being an hyperparameter to be determined from data ¹¹. It can be shown that the resulting likelihood of the targets, given the inputs, is

$$t \mid X, \theta, s_n \sim N(t \mid m, K + s_n^2 I_N)$$

with m and K being determined by the mean and covariance functions associated to the GP, evaluated at all the points in X .

For the posterior distribution, given all data $(x^{(i)}, t^{(i)})$, we direct the reader to our appendix on GPs, since it is of the same form as for the GP regression shown there.

Because the GP is not able to deal with high input dimensions($D \gg 20$), a technique for reducing the dimensionality of the input space is needed, and the authors decided to use the AS methodology.

To use it, we must assume the response function f can be well approximated by

$$f(x) \approx h(W^\top x)$$

where the matrix W is responsible for projecting the high-dimensional input x onto a lower-dimensional *active subspace*, and $h : \mathbb{R}^d \rightarrow \mathbb{R}$ is a lower-dimensional domain function known as a *link* function¹², with d being defined

¹¹Instead of using a full Bayes procedure, the authors decide simply to estimate the hyperparameters governing the GP prior by likelihood maximization. This perspective on estimation of hyperparameter is also known as an Empirical Bayes perspective

¹²It was not clear from the article, whether this assumption of univariate output was a limitation of AS or simply a matter of convenience of exposition

by the researcher¹³. This representation is not unique. It is usually assumed that W is orthonormal, living in what is called the Stiefel Manifold¹⁴.

To characterize the active space, the usual technique is to use the gradients $g^i = \nabla f(x^i)$.

If we define $C := \int \nabla f(x) (\nabla f(x))^\top \rho(x) dx$, with ρ being a density, then we can decompose the symmetric positive definite matrix $C = V\Lambda V^\top$, where $\Lambda = \text{Diag}(\lambda_i)$ - in decreasing order of magnitude $\lambda_i \leq \lambda_j$ for $i > j$. The usual approach proposes to use the first/largest d eigenvalues, associated to the directions of where the function variability is maximal, and so we can represent

$$V = \begin{bmatrix} V_d & V' \end{bmatrix}$$

and define

$$W = V_d$$

As usual, we cannot compute analytically C , and so we resort to using a Monte Carlo approximation

$$C_N = \frac{1}{N} \sum_{i=1}^N g^i (g^i)^\top$$

Now having found a W , we find the link function h by using a GP regression and treating the set $\{W^\top x^i\}$ as input points and the response function as h , and with the same target/output set t , i.e. $f(x) \approx h(W^\top x)$.

It still remains to explain how this AS and GP method can help us solve our optimization problem using a VFI procedure. After making a starting guess for the value function of V^∞ , at each iteration step s , we create n^s training input points in the state-space, $x_{1:n^s}^s$, then proceed to evaluate the Bellman operator on those points, obtaining several output points

$$t_{1:n^s}^s = \{t_1^s, \dots, t_{n^s}^s\}$$

$$t_i^s = TV^{s-1}(x_i^s).$$

Now we just apply the ASGP regression to TV^{s-1} as the response function, with input $x_{1:n^s}^s$, and output points $t_{1:n^s}^s$.¹⁵ And then, we solve the Bellman equation

$$TV^s(x^{s+1}) = \max_{\xi} \{r(x, \xi) + \beta E[TV^{s-1}(x^s)]\}$$

using the predictive mean of the resulting GP from TV^{s-1} .

¹³How to find a convenient d , the authors do not state.

¹⁴ $\{A \in \mathbb{R}^{D \times d} : A^\top A = I\}$

¹⁵To increase computational speed they parallelize this step.

In this setting, the authors restrict themselves to quantifying parametric and interpolation uncertainty obtained from ASGP. Here, the interpolation uncertainty originates from lack of available data since we are now running a surrogate model as a *global* solution of the economic model. Hence, they use the information about the predictive posterior to obtain some credible intervals.

Thus the authors decide to increase the state-space dimension by incorporating the parameter space, and assume that each parameter is independent with a uniform distribution. So, they learn a surrogate model from a space $\mathbb{R}^{\dim(x) \times \dim(\theta)}$, by also evaluating the Bellman operator on this extended space.

2.4 Uncertainty Quantification in Computer Models: Using GPs

In this section, we will look at a specific framework, which was created in the Computer Models subject with the objective of performing UQ, and in what way macroeconomists may use it for their research. A complete survey on the subject is clearly beyond the scope and purpose of this work, and the interested reader may find some good introductions to the subject in Soize (2018) and Sullivan (2015). For a better understanding of our work, and the inspiration for it, we decided to present the next framework, which accounts for several sources of uncertainty, with a greater analysis — specially on model bias — regarding its capabilities and limitations.

2.4.1 A Methodology for Recognizing Several Sources of Uncertainty

In Kennedy and O’Hagan (2001), a methodology is given for model uncertainty, even though the authors focus on modelling explicitly most sources of uncertainty, including code uncertainty, very similar to interpolation uncertainty. Code uncertainty results from too complex computer codes which cannot be run in efficient time, and for which we do not know their outputs at certain inputs of interest. Several examples of expensive computer codes are given such as the Gaussian Plume Model, used to predict the dispersion and subsequent deposition of radioactive material from an accidental release.

They formulate the relationship between the data observed $\{z_i : i =$

$1, \dots, n\}$, the true process $\zeta(\cdot)$ and the computer model output $\eta(\cdot, \cdot)$ by

$$\begin{aligned} z_i &= \zeta(x^i) + e_i \\ \zeta(x^i) &= \rho\eta(x^i, \theta) + \delta(x^i) \end{aligned}$$

where e_i is the measurement error for the i -th observation, ρ is an unknown regression parameter, and $\delta(\cdot)$ is the model discrepancy function/term which is modelled as being independent of $\eta(\cdot, \cdot)$. Let $y_i = \eta((x^i)^*, (\theta^i)^*)$ be the output of the computer model at some variable inputs $\{(x^i)^* : i = 1, \dots, N\}$ and parameter inputs $\{(\theta^i)^* : i = 1, \dots, N\}$. Both δ and η are given distinct GP as prior distributions, i.e. with different covariance functions c_η , c_δ , and mean functions

$$\begin{aligned} m_\eta(x, \theta) &= h_\eta(x, \theta)\beta_\eta \\ m_\delta(x, \theta) &= h_\delta(x, \theta)\beta_\delta \end{aligned}$$

with

$$p(\beta) = p(\beta_\eta, \beta_\delta) \propto 1$$

It would be interesting to now do a small hiatus in the exposition of the paper, and refer to Figure 2.3, in a section 2.1. By using a surrogate GP model on the computer model, we will be able to account for interpolation/code uncertainty, and parametric uncertainty, and by using a model bias/discrepancy term we will quantify for model/structural uncertainty.

Let us resume the exposition, defining $d = (y^\top, z^\top)^\top$, and supposing $e_i \stackrel{iid}{\sim} N(0, \lambda)$. Henceforth, to simplify the notation we will use $\phi = (\rho, \lambda, \psi)$, with ψ representing other hyperparameters which may be present in our covariance functions. Thus, we know that $d \mid \theta, \phi, \beta$, will be normally distributed.

To find the mean and variance, let us define, for computer outputs y , the following input points $D_\eta = \{((x^1)^*, (\theta^1)^*), \dots, ((x^N)^*, (\theta^N)^*)\}$, and, for observations z , the input points $D_\delta = \{x^1, \dots, x^n\}$, and $D_\delta(\theta) = \{(x^1, \theta^1), \dots, (x^n, \theta^n)\}$. Now writing

$$H_\eta(D_\eta) = \begin{pmatrix} h_\eta((x^1)^*, (\theta^1)^*) \\ \vdots \\ h_\eta((x^N)^*, (\theta^N)^*) \end{pmatrix}$$

and analogously for $H_\eta(D_\delta(\theta))$ and $H_\delta(D_\delta)$, we then get $E(y \mid \theta, \phi, \beta) = H_\eta(D_\eta)\beta_\eta$, and similarly $E(z \mid \theta, \phi, \beta) = \rho H_\eta(D_\delta(\theta))\beta_\eta + H_\delta(D_\delta)\beta_\delta$ ¹⁶ and

$$m_d(\theta) := E(d \mid \beta) = H(\theta)\beta$$

¹⁶In these expectations, we are conditioning also on the input variables.

where

$$H(\theta) = \begin{pmatrix} H_\eta(D_\eta) & 0 \\ \rho H_\eta(D_\delta(\theta)) & H_\delta(D_\delta) \end{pmatrix}$$

To formulate $\text{var}(d \mid \theta, \beta, \phi)$, we first need to define the variance matrix

$$\text{var}(y \mid \theta, \beta, \phi) = V_\eta(D_\eta) = \left[c_\eta(((x^i)^*, (\theta^i)^*), ((x^j)^*, (\theta^j)^*)) \right]_{i,j=1,\dots,N}$$

Defining in an analogous manner $V_\eta(D_\delta(\theta))$ and $V_\delta(D_\delta)$ and

$$C_\eta(D_\eta, D_\delta(\theta)) = \left[c_\eta(((x^i)^*, (\theta^i)^*), (x^j, \theta^j)) \right]_{i=1,\dots,N; j=1,\dots,n}$$

and so, keeping in mind that the GP are assumed independent of each other, it can be shown that

$$\text{var}(d \mid \theta, \beta, \phi) = V_d(\theta) = \begin{pmatrix} V_\eta(D_\eta) & \rho C_\eta(D_\eta, D_\delta(\theta))^\top \\ \rho C_\eta(D_\eta, D_\delta(\theta)) & \lambda I_n + \rho^2 V_\eta(D_\delta(\theta)) + V_\delta(D_\delta) \end{pmatrix}$$

Therefore, using as prior distribution

$$p(\theta, \beta, \phi) = p(\theta)p(\phi)$$

and since we have just deduced the distribution of d , and hence also the likelihood, the authors obtain the following full joint posterior distribution

$$p(\theta, \beta, \phi \mid d) \propto p(\theta)p(\phi)N(d \mid m_d(\theta), V_d(\theta))$$

Estimating hyperparameters Due to the numerical intractability of the resulting expression with respect to ϕ ¹⁷, after marginalizing out β from the previous joint posterior distribution, making impossible a full Bayesian analysis, the authors propose instead to fix the hyperparameters ϕ at a reasonable estimations. Therefore, for inference on θ , its conditional posterior given fixed values of ϕ will be used. This process can be divided in two stages:

- First Stage - We use $\{y^i\}$ (computer output only) to estimate the hyperparameters of the GP governing η , namely β_η , those of h_η , and of c_η .
- Second Stage - Considering fixed the hyperparameters for η , and using $\{z^i\}$ (the field/experiment observations), we estimate ρ , λ and the hyperparameters of the GP governing δ .

¹⁷ ϕ includes the parameters of the GP covariance and mean function, which are usually at least 6, making intractable numerical integration

The cost of using this modularization is that some information on η in $\{z^i\}$ may be lost, and hence this method does not account fully for all sources of uncertainty. One cannot be sure about what effects in UQ will not be considered, but Kennedy and O'Hagan (2001) state that according to their experiments the loss can be negligible. However, no evidence or experiment is presented to show the reach of such claim.

Prediction of ζ The posterior distribution of ζ conditional on the hyperparameters of the GPs, and of the calibration inputs θ is also a Gaussian Process, with a mean function¹⁸

$$E(z(x) \mid \theta, \phi, d) = h(x, \theta)^\top \hat{\beta}(\theta) + K(x, \theta)^\top V_d(\theta)^{-1} (d - H(\theta) \hat{\beta}(\theta))$$

where

$$h(x, \theta) = \begin{pmatrix} \rho h_\eta(x, \theta) \\ h_\delta(x) \end{pmatrix}$$

and

$$K(x, \theta) = \begin{pmatrix} \rho C_\eta((x, \theta), D_\eta) \\ \rho^2 C_\eta((x, \theta), D_\delta(\theta)) + C_\delta(x, D_\delta) \end{pmatrix}$$

and a covariance function

$$\begin{aligned} \text{cov}(\zeta(x), \zeta(x')) &= K(x, x') \\ &= \rho^2 c_\eta((x, \theta), (x', \theta)) + c_\delta(x, x') - K(x, \theta)^\top V_d(\theta)^{-1} K(x', \theta) \\ &\quad + \left(h(x, \theta) - H(\theta)^\top V_d(\theta)^{-1} K(x, \theta) \right)^\top \\ &\quad W(\theta) \left(h(x', \theta) - H(\theta)^\top V_d(\theta)^{-1} K(x', \theta) \right) \end{aligned}$$

with $W(\theta) = (H(\theta)^\top V_d(\theta)^{-1} H(\theta))^{-1}$. Hence, using the distribution of $\zeta(x) \mid \theta, \phi, d$ and together with that of $\theta \mid \phi, d$, we can use numerical or sampling methods to characterize the distribution of $\zeta(x) \mid \hat{\phi}, d$, where $\hat{\phi}$ are the fixed estimates for the hyperparameters of the GPs.

For an uncertainty analysis such as parametric variability on the input variables, the authors consider imposing a distribution on the input variables, G_X . Now, quantifying uncertainty such as interpolation, or parametric, is to infer on the distribution of $\zeta(X)$. And so, they obtain $E_X(\zeta(X)) = \int \zeta(x) dG_X(x)$, and similarly for other quantities, all derived from the posterior of ζ .

¹⁸These formulas come from a GP regression

2.4.2 Identifiability Issues and Possible Solutions

The method proposed by Kennedy and O’Hagan (2001) raises some interesting questions on possible confounding issues. Most methods accounting for the sources of uncertainty treat the scientific model as a black box. We will focus, in our work, on those methods which respect the black-box treatment of the computer model. An example of a method which uses specific knowledge from the scientific model to improve identifiability of the calibration parameters can be found in V. R. Joseph and Yan (2015).

At this point, it would be useful to differentiate the different *possible* identification issues. For that purpose, and to facilitate the reader’s work, we rewrite the following equation:

$$\begin{aligned} z_i &= \zeta(x^i) + e_i \\ \zeta(x^i) &= \rho\eta(x^i, \theta) + \delta(x^i) \end{aligned}$$

Confounding of model discrepancy with measurement error It is analogous to the case of endogeneity, and where one would use an IV methodology (see for example Angrist and Krueger (2001)). In Kennedy and O’Hagan (2001), it was assumed that errors were independent from the input variables. Why this would be the case deserves some extra consideration. Our true model ζ acts as a GP, i.e. a non-parametric model capturing most of the dynamics of the input variables, if not all. As a whole, ζ will not lack identification, but it may not be clear which term, whether the surrogate model or the discrepancy term would capture part of the behaviour of the error term. From how the hyperparameters were estimated in the paper, it is the discrepancy term, fitted to the field data, which will learn most of the dynamics. In this regard, the discrepancy term will behave in a certain way as an IV (Instrumental Variable), removing any possible endogeneity¹⁹ between the input variables, through the computer model term, and the error term.

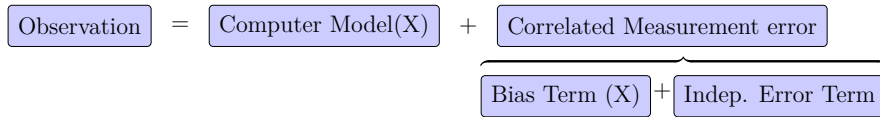


Figure 2.2: Endogeneity when not accounting for bias term.

Several other methods have also been used, for example *panel data* methods (see Arellano (2003)), or *fixed* and *random effects* models. It is interesting

¹⁹Endogeneity can sometimes be seen as a special case of unobserved heterogeneity. Other authors, identify the two.

that since the addition of the discrepancy term could be interpreted in a way as adding a random effect, it will also share some of its identifiability issues, as we will see later. Furthermore, some other care must be taken, when using Kennedy and O'Hagan method, since we also have parameters ρ , from the computer model term, and λ , the error term variance, being learnt from field data. As long as these behave as constants w.r.t. to input variables, the discrepancy term will capture most of the dynamics in the measurement error. However, if we allow the ρ to capture some dynamic in the field data, then we may have an identifiability issue. The authors, apparently oblivious to this possibility, state, in their motivation for the definition of ζ , that ρ «may formally depend on x ». Although the presence of this confounding may be inescapable in Economics, a subject where experiments/data collection is very complex, the method proposed by the authors deals with unobserved heterogeneity in specific contexts, and still complies with the purpose of UQ, i.e. to account and measure sources of uncertainty, even though there might be some confounding between the effect of the model bias term being due to measurement uncertainty or due to model uncertainty.

Confounding between Computer/Surrogate Model and Discrepancy Term In the discussion following Kennedy and O'Hagan (2001), this issue was also raised by H. P. Wynn²⁰, and the reply by the authors touched only lightly the issue of the previous paragraph, ignoring(or confounding?) completely this other topic. In M. J. Bayarri et al. (2007), a purely theoretical toy example is given on how there could be confounding between the computer model term and the bias term. An intuitive explanation for the issue is when the same predicted experimental response can result from many different combinations of parameter inputs and the model discrepancy function.

This explains why we may have an accurate prediction even in the presence of non-identification of the calibration parameters. Some possible symptoms of such a challenging context, include a large posterior variance of the respective calibration parameter, together with a large width in the prediction intervals of the discrepancy function, which will for a wide range of different values for parameter inputs and the model discrepancy function. Moreover, having access to more observations/data will not alleviate the problem.

One method to handle this issue is using Bayesian analysis, by considering

²⁰Just think of two people on the same scale/weighing machine. I can estimate the group of two people, but it seems impossible to estimate correctly the weight of each person. This example was also authored by H. P. Wynn in that same discussion.

a careful choice of priors in such a way that leads to identification. Of course this is not innocuous, for when doing extrapolation, the priors may be highly influential in the predictive performance. In this regard, some expert knowledge, as always when eliciting priors, may be of help, but it may not be clear in what manner this could infringe on our assumption of treating the computer model as a black box. However, even if we were to simply disregard the black box consideration, the fact that the calibration parameters we are usually interested in estimating are unknown, may make the elicitation of useful priors much harder, if not impossible.

Similarly to the consultation of expert knowledge for the elicitation of tighter priors around important parameter values, we could also consider the smoothness of the GP, determined by its covariance function. Smoother processes, will tighten the posterior distributions of the parameters, indicating the possibility of identification. An intuition for this, using the weighing machine example, is to think as having the posterior distributions for the weights of two people with very thick tails. We can have a total of 160kg with (10,150) or (150,10) combinations. If we know that both persons are fit adults, then we can choose tight priors for both, around 80kg. For more details, the interested reader may check Arendt, Apley, and Chen (2012).

For another method which may help us solve this confounding, observe how the hyperparameters were estimated in Kennedy and O’Hagan (2001). The authors put forward reasons of tractability and dimensionality to an empirical Bayes analysis, where computer output was used to identify the hyperparameters of the surrogate process, and field data was used to identify the discrepancy term, instead of doing a full Bayesian analysis. Unbeknownst to them, this separation of hyperparameters identification became later known as *modularization* in Bayesian analysis. The reference F. Liu, M. Bayarri, and Berger (2009) studies the topic of modularization in greater detail, and puts forth several reasons why it may be advisable to use modularization when doing a Bayesian analysis of computer models. Among them, we find an improved mixing of MCMC draws, tractability, and most importantly « identifiability concerns or confounding ». Similarly, Arendt, Apley, and Chen (2012) besides exemplifying with a real engineering model the problem of non-identification, the authors also investigate specifically how modularization may help identify the Gaussian Processes in question. Furthermore, they also stress the confounding issue of the calibration parameters θ , present in the mean of the GP, and the discrepancy term. We will tackle this other issue in a few paragraphs below.

This proposed solution of modularization, as previously stated, does restrict our accounting of sources of uncertainty, and in this regard it is not exactly clear what type of dynamics are being lost when we proceed with

modularization.

In reference Arendt, Apley, Chen, et al. (2012), a continuation of the research developed in Arendt, Apley, and Chen (2012), the authors show how identifiability can be improved upon by using multiple output/responses, all dependent on the *equal* set of parameter and variable inputs. In Jiang et al. (2017) a formal methodology, denominated Preposterior Analyses Method, is suggested for choosing a subset of responses, from a multiple output modelling, for enhancing identifiability.

Confounding in a non-zero mean GP The confounding present when the mean function of the GP has unknown coefficients was not explicitly recognized either by the authors or by the discussants of the Kennedy and O’Hagan (2001) paper. Perhaps because when we have a deterministic computer model term without the need for a probabilistic surrogate model, we could see ζ as non-zero mean GP. However, this type of confounding may also be encountered when we must assume either the discrepancy or the surrogate model to be a non-zero mean GP. In the relatively recent reference Plumlee and V. Joseph (2018), we can find an example where such identifiability problem is troublesome, leading to an unreasonable estimate of the mean function. However, prediction of the quantities of interest are still very good, despite the interpretability of the mean function being lost due to this issue.

In fact, this identifiability issue is in a way shared by a more common modelling technique in Spatial Statistics called *spatial random effects modelling*. This issue was first formally recognized in Reich and Hodges (2010) and Reich, Hodges, and Zadnik (2006), where it is shown that the introduction of a spatial structure by the inclusion of a random effect would induce a significant change in the posterior estimate of a fixed effect, contrary to what was quite plain from non-spatial analysis. This problem originates from the characteristics of data, since sites with similar covariates to neighbouring sites cannot help distinguish between the competing explanations of spatial clustering and the fixed effects, since both will predict similar behaviours for them and the neighbouring sites. This may be interpreted as akin to the effect of adding a collinear regressor to a linear model. For a more formal and systematic listing of the conditions for confounding when we add a spatial random effect, see Reich and Hodges (2010). However, when neighbour sites have very different covariates, we can have information that will help us distinguish from both possible effects, which may allow for the preservation of the fixed effect estimation, with or without the inclusion of spatial structure. To be sure that this is a case of spatial confounding, the authors propose

a restricted spatial regression, which is characterised by restricting the spatial random effect to a subspace orthogonal (and hence not correlated) to the fixed effect covariates. When using this orthogonal regression, the original non-null fixed effect of the non-spatial analysis is salvaged.

In reference Plumlee and V. Joseph (2018), we can also find a proposed resolution to the identifiability problem in GP regression, inspired in the restricted spatial regression above. The method suggests constructing a covariance function in order to orthogonalise the Gaussian Process term with respect to the mean function. However, some assumptions seem very restrictive, such as the mean function being of the form $m(x) = \beta^\top g(x)$, where we are interested in estimating the β , and $g(\cdot)$ must be a *known* function. Although for emulation purposes, it may not be very restrictive, for the context of our work, where the mean function will be the economic model at hand, which cannot not be written as a polynomial in general, it does seem to be. In Chapter 4, this will become clearer. Furthermore, the constructive procedure for the desired covariance function requires the use of numerical approximations, adding more pressure to the computational performance of the overall method.

Although the identifiability issues may be a serious concern in some contexts, namely in preventing from estimating the true parameter values, as stated in Arendt, Apley, and Chen (2012), research has been done showing that predictive accuracy is not affected when either confounding issues are present. See for example Jiang et al. (2017), Loeppky, Bingham, and Welch (2006), and Plumlee and V. Joseph (2018) and references therein.

2.5 Contributions of Our Work

«Models are to be used, not believed»

Henri Theil, Principles of Econometrics ²¹.

Our work will focus on the topic of model uncertainty in macroeconomics. As in any scientific subject, the macroeconomic models used are abstractions, simplifications of reality, of the true underlying process, which is usually, if not always, unknown to us. The Data Generation Process (DGP), for the purpose of our study, will be represented by the reality that we have access to, at any given moment.

²¹As cited in DeJong and Dave (2011)

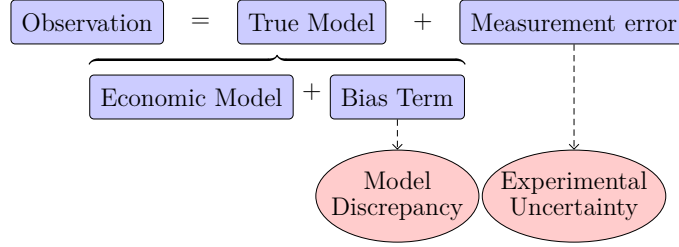


Figure 2.3: Our approach to UQ in this work

In our work, and in Economics at large, the DGP will no longer be a physical, chemical or biological experiment, but society, or Man himself. This makes our task of quantifying model bias potentially much harder due to the impossibility to conduct certain experiments on our object of study, or simply the sheer complexity of collecting data at a macro level from a human 'sensor'. This makes our modelling of measurement error much more difficult, and hence our bias term may easily capture also other sources of uncertainty. Our framework is not without limitations, almost surely will confound measurement and model bias, and there may be other identification issues as the ones stated previously, but without it, consequences of not accounting for these sources of uncertainty could be much more serious. This thesis is a contribution to a more proper and more complete UQ analysis in macroeconomics.

One common symptom of this structural uncertainty is the problem of parameter overfitting, i.e. the best fit value of the parameter may be inconsistent with its scientific interpretation. This usually occurs when the parameters try to compensate for the models limitations, and hence take values for which our theory has some difficulties to harmonize with.

The concept of calibration in Macroeconomics took the very specific meaning of the act of choosing values for the model's parameters based on microeconomic studies, or 'known economic facts'²². This method of calibration became prevalent in the subject of Macroeconomics, specially when dealing with RBC models, since by reducing the parameter space, better estimations for the remaining free parameters were often found. The curious reader is directed to DeJong and Dave (2011). This calibration methodology may be seen as a procedure to deal with the overfitting problem, but it certainly does not tackle what may be the root of the problem, model discrepancy.

²²This is very different to what may be understood as calibration in the area of Bayesian Inference(fitting), Machine Learning(transforming class scores into class membership probabilities), or in Statistics at large (known dependent variable observations are used to predict corresponding explanatory variables).

So far, most methods from the formal subject UQ used in Macroeconomics have been centered on parametric uncertainty propagation. In our work, we focus on model uncertainty, or more precisely on misspecification. In the next chapter, we present a literature review on the most usual and recent methods used in Macroeconomics to deal with misspecification.

One of our contributions to this research area is the introduction of a non-parametric Bayesian method, able to account for very complex non-linear dynamics, with some interesting results when compared to a more usual method. Because our method treats the scientific/computer model as a black box, it may be adapted to most economic models, linear or non-linear.

The main objective of our work will be to increase awareness in the macroeconomics subject to this new area of Uncertainty Quantification, by employing methods from the latter that may help reduce and account for model discrepancy/misspecification, meanwhile improving the models' forecast performance, and potentially maintaining the economic interpretation of parameter values.

Chapter 3

Dealing with Misspecification in Macroeconomics

This chapter serves as a survey on the several methods researchers have available to assess Dynamic Stochastic General Equilibrium (DSGE) models misspecification and possibly correcting it. The objective is to illustrate how most methods either assume the model must be linear, or the type of dynamics that are able to capture, and correcting, may pale when compared to what the methodology proposed in this work can achieve.

3.1 Dynamic Stochastic General Equilibrium Models (DSGE)

In our work, we will focus on the misspecification of DSGE models. Below is an explanation of the designation for this class of models.

Dynamic Intertemporal optimization

Stochastic Optimization of a stochastic objective function

General Equilibrium The modelling is done when all markets (goods and services, labour, financial, etc.) are in equilibrium.

As was stated in Chapter 2, after the financial crisis, there has been a renewed effort to endow the DSGE models with the ability to capture ever more complex dynamics. However, regardless of these or future improvements, the scientific model will never be able to describe such a complex

phenomenon as is reality, and hence, for those who have to provide recommendations on policy decisions, instead of ignoring the misspecification of models and naively use the DSGE as the true model, one may try to account for model misspecification.

A DSGE model, after being linearised at equilibrium around the steady-state, may have a state-space representation of the following form

$$\begin{cases} z_t = \mathbf{H}(\theta)s_t + v_t, & \text{Measurement Equation} \\ s_t = \mathbf{B}(\theta)s_{t-1} + w_t, & \text{State Transition Equation} \end{cases}$$

where θ is the parameter vector of the linearized model, z_t are the observed variables, and s_t are the latent/state variables. Although we concern ourselves with a linear model, our method could also be adapted to non-linear one. However, one assumption we do make is to work with matrices \mathbf{H} and \mathbf{B} which are time-invariant, to ensure tractability. For non-linear models, terms of the form $\mathbf{B}_2(\theta)(s \otimes s)$, or of greater degree than 2, would show, but we must still assume that \mathbf{B}_2 is time-invariant for our proposed method to be used.

There are several possible sources of model uncertainty in a DSGE model, namely we could have insufficiently complex non-linear dynamics, either from working from an approximation of the theoretical model or from the theoretical model itself, we could also have neglected important variables leading to missing channels. In fact, many DSGE models up to 2007 lacked a financial sector, which given this last financial crisis, makes this criticism gain some weight. Another possibility maybe some of the assumptions on the shock structure of theoretical DSGE, i.e. the shocks might not be well modelled as an AR(1) process.

Hence, a key part in evaluating DSGE prediction performance is also the recognition of model misspecification, and possible measures to deal with it. It is foreseeable that a better specified model will result in improved performance, although that may not necessarily mean an identified model. In fact, the blind pursuit of identification, disregarding whether those extra assumed assumptions guaranteeing identification are correct, may be counter-productive, at least policy wise, as was exemplified by Kocherlakota (2007). In this interesting article, it is shown that even a perfectly fitting model may provide more misleading answers to policy questions than a less misspecified model.¹

The benchmark model against which many of the published DSGE models are compared to is that of Smets and Wouters (2007). This medium-sized

¹From Kocherlakota (2007): «Which one works better depends on which incorrect assumption is a better approximation to reality»

DSGE model (SW model) was first developed in Smets and Wouters (2003), where several features such as capital (with costs of adjustment), investment, habit formation in consumption, and variable capacity utilization were introduced, while maintaining a New-Keynesian setting of sticky prices and wages for the Euro Area. In Smets and Wouters (2007), the same baseline model is used but with three main differences. Firstly, the number of shocks is reduced to the number of observables. Secondly, the model includes a deterministic growth rate resulting from labour-augmenting technological progress, which avoids the need for detrending the data before estimation. Thirdly, the Kimball aggregator is used in the intermediate goods and labour market instead of the Dixit-Stiglitz one. This substitution allows them to estimate a more reasonable degree of wage and price stickiness. In Poudyal and Spanos (2016)², the SW model from Smets and Wouters (2007) is tested and shown to be misspecified. In fact, the implicit resulting statistical model from the SW model is a Normal VAR, but when the authors try to correct the misspecification, they arrive at a Student's-t VAR model.

3.2 Hybrid Models

The Hybrid Models approach is one of the most disseminated approaches to misspecification correction in the Macroeconomics literature. It is characterised by a relaxation of the constraints on the DSGE model, which can be broadly classified into two methods: Additive Hybrid Models, and Hierarchical Hybrid Models. The classification presented in this section of Hybrid Models will follow closely that of Paccagnini (2017), which is a survey on this literature, where the interested reader may find more information on the subject of misspecification in macroeconomic structural modelling.

3.2.1 Additive Hybrid Models

This methodology simply takes the original SSM and modifies the dynamics of a component appearing as an additive term in the measurement equation. Some more complex methods may add some extra relationships to the SSM as will see below. To the best of our knowledge, so far, all methods deal within the framework of linear State-Space model, and their modifications capture again only linear dynamics.

²Only the 2016 version of this paper assesses the SW model.

Additive Model à la Ireland

$$\begin{aligned} z_t &= \mathbf{H}(\theta)s_t + \Lambda_0 + \Lambda_1 t + \Lambda_u u_t \\ s_t &= \mathbf{B}(\theta)s_{t-1} + w_t \end{aligned}$$

where $u_t = \Gamma_1 u_{t-1} + \Gamma_\epsilon \epsilon_t$ is a VAR process and may be interpreted as a measurement error.

This model was first presented in Ireland (2004), using as a framework an RBC model, where Γ_1 , and Γ_ϵ were allowed to have non-zero, off-diagonal entries. This method will also be further explored later on, in Chapter 4, when we compare our methodology's results to those in Ireland (2004).

Some known advantages on the inclusion of this VAR term are the fit improvement of the estimated augmented/empirical model to the data, with potentially better parameter estimates. On the negative side, its structure may be just too simple to truly represent the measurement error and model misspecification.

Additive Model with a Dynamic Factor Equation (DSGE-DFM)

In Guerron-Quintana (2010), we learn just how sensitive to the exact data of observables the estimation of structural parameters in DSGE modelling may be. Sometimes, researchers are in a data-rich environment, i.e. they also have access to data which is not specified explicitly in the DSGE model. In Boivin and Giannoni (2006), besides having the usual additive hybrid model, the authors add an extra equation connecting a large vector of non-modelled variables, similar to a Dynamic Factor Model (DFM)

$$\begin{aligned} z_t &= \mathbf{H}(\theta)s_t + \Lambda_{z0} + \Lambda_{z1}t + u_t \\ s_t &= \mathbf{B}(\theta)s_{t-1} + w_{s,t} \\ F_t &= \Lambda_{F0} + \Lambda_{F1}t + \Lambda_s s_t + w_{F,t} \end{aligned}$$

where u_t (VAR), $w_{s,t}$ and $w_{F,t}$ are uncorrelated.

There are a few other examples of additive hybrid modelling that have showed little improvement over the presented above models, such as in Schorfheide, Sill, and Kryskho (2010) with a simplified version of the model above, and also Canova (2014) where it is suggested a method to capture the long-run features of data, which usual DSGE models have difficulties in. Regardless, they still assume a linear SSM as their framework.

3.2.2 Hierarchical Hybrid Models

In the hierarchical hybrid models, we use perturbed versions of the matrices determined in our DSGE modelling.

$$\begin{aligned} z_t &= \mathbf{H}_1(\theta)s_t + \mathbf{H}_0 + u_t \\ s_t &= \mathbf{B}_1(\theta)s_{t-1} + \mathbf{B}_0 + \mathbf{B}_w w_{s,t} \end{aligned}$$

where we now interpret \mathbf{B}_i and \mathbf{H}_i as perturbed versions of the original matrices by disturbances η_i^B and η_i^H , related by

$$\begin{aligned} \mathbf{H}_i &= \Phi_i(\theta) + \eta_i^H \\ \mathbf{B}_j &= \Psi_j(\theta) + \eta_j^B \\ i &\in \{0, 1\}, j \in \{0, 1, w\}, \end{aligned}$$

Again, all these models assume specifically to be dealing with a linear SSM, to the best of our knowledge.

The DSGE-VAR of Del Negro and Schorfheide

Based on the work of B. Ingram and Whiteman (1994), this method of creating a DSGE-VAR was introduced in Negro and Schorfheide (2004) as a tool to improve forecasting for misspecified models, and was soon extended to policy evaluation in Negro and Schorfheide (2009).

This method of model assessment, if applied to a VARMA representation instead to a VAR, would increase significantly the complexity of the posterior computations, and since a VAR with sufficient lags can be shown, in many cases, to approximate well a DSGE model, the method is broadly applicable.

To find a VAR approximation to a DSGE model, i.e. a DSGE-VAR, let's consider the following VAR representation:

$$z_t = A_0 + A_1 z_{t-1} + \dots + A_p z_{t-p} + u_t, \quad E[u_t u_t'] = \Sigma$$

Defining $X_t = (1, z_{t-1}', \dots, z_{t-p}')'$, and $A = (A_0, \dots, A_p)'$, we can write the VAR with p-lags in the companion form $Z = XA + U$.

The usual VAR moment restrictions apply:

$$A(\theta) = E_\theta[X_t X_t']^{-1} E_\theta[X_t z_t'] \quad \Sigma(\theta) = E_\theta[z_t z_t'] - E_\theta[z_t X_t'] E_\theta[X_t X_t']^{-1} E_\theta[X_t z_t']$$

With this method, we must consider two sources of misspecification. The approximation by the VAR to the DSGE³, which could be in theory be minimized by a sufficiently good approximation, and the DSGE model itself may be misspecified.

With the purpose of capturing the misspecification from the DSGE, the authors introduce matrices A^δ and Σ^δ in

$$A^{miss} = A(\theta) + A^\delta, \quad \Sigma^{miss} = \Sigma(\theta) + \Sigma^\delta$$

where

$$A^{miss} | \Sigma^{miss}, \theta \sim N \left(A(\theta), \frac{1}{\lambda T} ((\Sigma^{miss})^{-1} \otimes E_\theta[X_t X_t'])^{-1} \right)$$

and

$$\Sigma^{miss} | \theta \sim \text{Inv-Wish}(\lambda T \Sigma(\theta), \lambda T - k, n)$$

has an Inverse-Wishart prior distribution.

The hyperparameter λ scales the precision of the prior. The smaller the values of λ , the more diffuse is the prior. In the limit, the hyperparameter will shift all prior weight to the VAR approximation of the DSGE model. These priors are conjugate, and hence the posterior will belong to the same probability distribution family. The role for λ will be similar in the posterior, since the higher its value, the larger the weight put on the moment restrictions for the VAR approximation of the DSGE model.

The question of which value should be chosen for the hyperparameter is usually data-driven, the hallmark of a Bayesian analysis. Defining the marginal data density as

$$p_\lambda(Y) = \int p_\lambda(Y) p(\theta) d\theta$$

It's usual to restrict the possible values that λ can take to a finite grid Λ , and if one assumes an equal probability for each point of the grid, $p(\lambda|Y) \propto p_\lambda(Y)$, i.e. we can consider the posterior density of λ as proportional to the marginal data density. The assessment of the model can be done by considering the posterior distribution of the hyperparameter. Higher mass on the higher values of λ are evidence in favor of the VAR approximation to the DSGE, and hence of the DSGE.

A simple estimator of the hyperparameter would be

$$\hat{\lambda} = \underset{\lambda \in \Lambda}{\operatorname{argmin}} p_\lambda(Y)$$

³A DSGE model can usually be shown to have a VARMA representation. For example, the model in Smets and Wouters (2007), the SW model, was shown to have a VARMA(2,1) representation in Morris (2016)

As a matter of convention, intermediate values, where with the method developed so far the data is not explicit on the (in)validity of the model, are considered to be between 0.5 and 2. In those situations, it is suggested to compare the impulse response functions between the DSGE-VAR($\hat{\lambda}$) and the DSGE-VAR(∞).

Since the DSGE-VAR(∞), obtained from the moment conditions, was given in reduced form, and for the IRFs we need to identify the structural shocks, the authors represent the disturbance errors u_t ⁴ from the structural disturbance errors ϵ_t by

$$u_t = \Sigma_{chol}^{miss} \Omega \epsilon_t$$

where Σ_{chol}^{miss} is the Cholesky decomposition of Σ^{miss} and Ω is an orthonormal matrix which is not identified from the assumptions presented so far. In Negro and Schorfheide (2004), a method is proposed to find such a matrix. In the DSGE-VAR(∞)⁵, the first impact of a structural shock in the z_t vector, $B(\theta)$, can be decomposed uniquely⁶ into

$$B(\theta) = A(\theta)\Omega(\theta).$$

One can also determine a similar equality for the other DSGE-VAR,

$$B(\theta)^{miss} = A^{miss}\Omega^{miss}$$

where the authors propose $\Omega^{miss} := \Omega(\theta)$. We can interpret this suggestion as matching the dynamics of the DSGE-VAR IRF to those of the DSGE-VAR(∞), so in the case of no misspecification both models will have the same IRFs. This paper also gives an MCMC algorithm, based on a random walk MH proposal step, to estimate the DSGE-VAR(λ).

The DSGE-FAVAR

We previously stated that the VAR approximation could be a source of misspecification. In Consolo, Favero, and Paccagnini (2009), several symptoms of possible existing misspecification from that approximation are given, one of them being the instability of the reduced form VARs estimated parameters. This is against what one would expect, since those parameters are functions of the structural parameters which are constant over time. With the possibility of losing important information by using a DSGE-VAR representation, the authors of Consolo, Favero, and Paccagnini (2009) create a

⁴The disturbances are also the one-step-ahead forecast errors

⁵The VAR approximation to the DSGE model

⁶Using QR decomposition

DSGE-FAVAR with the following specification:

$$\begin{pmatrix} Z_t \\ F_t \end{pmatrix} = \begin{pmatrix} \Phi_{11}(L) & \Phi_{12}(L) \\ \Phi_{21}(L) & \Phi_{22}(L) \end{pmatrix} \begin{pmatrix} Z_{t-1} \\ F_{t-1} \end{pmatrix} + \begin{pmatrix} u_t^Z \\ u_t^F \end{pmatrix}$$

where Z_t are observable variables specified in the DSGE model, and F_t are unmodelled factors obtained from an additional macroeconomic time-series considered relevant to Z_t . This method is then implemented in an analogous way to the DSGE-VAR explained above. For more information, beyond the scope of this work, see Consolo, Favero, and Paccagnini (2009).

3.3 Bayesian Model Averaging

Usually Bayesian Model Averaging (BMA) is associated to the task of model selection, and can also be used as method to account for model uncertainty.

For each model M_i , $i = 1, \dots, I$, under consideration, let us assign a prior probability $\pi(M_i)$. Each model M_i has a corresponding set of parameters θ_i , which in turn also has an associated distribution given the model, i.e. $\pi(\theta_i | M_i)$.

Given a model, we can also obtain the posterior distribution of each θ_i , which is given by

$$\pi(\theta_i | Y, M_i) = \frac{L(Y | \theta_i, M_i) \pi(\theta_i | M_i)}{\int L(Y | \theta_i, M_i) \pi(\theta_i | M_i) d\theta_i}.$$

where Y is the data, and $L(Y | \theta_i, M_i)$ is the likelihood function conditioned on model M_i . The quantity $\pi(Y | M_i) = \int L(Y | \theta_i, M_i) \pi(\theta_i | M_i) d\theta_i$ is called the model's marginal likelihood and plays an essential part in BMA.

From Bayes' theorem we have

$$\pi(M_i | Y) = \frac{\pi(Y | M_i) \pi(M_i)}{\sum_{i=1}^I \pi(Y | M_i) \pi(M_i)}.$$

These probabilities can be used for a model selection criteria, and have also a connection to what is called Bayes factor of model i against m , which defined by

$$\text{BF}_{im} = \frac{\pi(M_i | Y)}{\pi(M_m | Y)}.$$

We can rewrite the posterior distribution of model M_i with respect to Bayes factors by considering one of the models as the baseline, which we will denote by M_1 , we obtain

$$\pi(M_i | Y) = \frac{\text{BF}_{i1} \pi(M_i)}{\sum_{i=1}^I \text{BF}_{i1} \pi(M_i)}.$$

Another use of BMA, besides model selection, is to account for model uncertainty. Considering a QoI z , obtainable from all models, its marginal posterior distribution across all models is

$$\pi(z | Y) = \sum_{i=1}^I \pi(z | Y, M_i) \pi(M_i | Y) .$$

This way, we are considering the contribution of all models to the quantity of interest, according to each model’s prior distribution and allowing any of them to be potentially the true model. For further information, the interested reader is directed to Fragoso, Bertoli, and Louzada (2018) and Steel (2019).

One of the major limitations of BMA lies precisely in allowing any of the models to be considered correct. What will happen when none of the models is the true model, and their performance for the quantity of interest is poor? In this context, BMA does not seem to allow for a proper misspecification correction. In Clarke (2003), instead of the BMA converging to an element of the closed convex hull of the model list, it may converge only to the single best element of the model list, overfitting the data.

3.4 Recent Advances

3.4.1 Identification of misspecification in Data-Rich Environments

In Monti (2015), a method is proposed to include observables, other than those which the model in its state-representation already does, in an attempt to identify which of them could have a Granger-causality type of relationship with the state variables. The reasoning behind this is if a model is well specified, then no ‘un-modelled’ variable should improve prediction of the QoI. If the reasearcher can conclude for Granger causality, then those extra observables may indicate a direction in which the model should be improved.

The gist of the method is to jointly model the states of the DSGE and the auxiliary/extra variables as a Bayesian VAR (BVAR). After obtaining the posterior distribution for the parameters, the researcher verifies which coefficients are different from zero, i.e. which auxiliary variable is Granger-causing the states of the DSGE model, by checking for which posterior a 90% credible interval does not include zero. They also do a Forecast Error Variance Decomposition (FEVD) analysis to verify how much weight is given to the ‘un-modelled’ variables.

After solving and linearising the DSGE, we will have the following SSM:

$$\begin{aligned} z_t &= \mathbf{H}(\theta)s_t + \xi_t \\ s_t &= \mathbf{B}_1(\theta)s_t + \mathbf{B}_0(\theta)\epsilon_t \end{aligned}$$

Let us denote the extra variables by $X_t = (x_{1,t}, \dots, x_{N,t})^\top$. We will model X_t as

$$X_t = \Pi_1 X_{t-1} + \dots + \Pi_p X_{t-p} + \Gamma_0 \mathbf{B}_1 s_{t-1} + \Gamma_0 \mathbf{B}_0 \epsilon_t + \xi_t .$$

This BVARX(p) model is estimated using a modified Litterman prior. Then, by treating the states as observables,

$$\begin{bmatrix} s_t \\ \tilde{X}_t \end{bmatrix} = \begin{bmatrix} \mathbf{B}_1 & \mathbf{0} \\ \Gamma_0 \mathbf{B}_1 & \Pi \end{bmatrix} \begin{bmatrix} s_{t-1} \\ \tilde{X}_{t-1} \end{bmatrix} + \begin{bmatrix} I & \mathbf{0} \\ \Gamma_0 \mathbf{B}_0 & V \end{bmatrix} \begin{bmatrix} \epsilon_t \\ \xi_t \end{bmatrix}$$

where $\tilde{X}_t = [X_t^\top X_{t-1}^\top \dots X_{t-p+1}^\top]^\top$.

This last model is to be used as a prior for the joint time series $y_t = [s_t^\top \tilde{X}_t^\top]^\top$. It is important to notice that the prior is conservative in the sense that it imposes no Granger-causality, which can be seen by the null submatrice.

The process governing y_t is assumed to be

$$\begin{bmatrix} s_t \\ \tilde{X}_t \end{bmatrix} = \begin{bmatrix} B(\theta) & B_{21} \\ \Gamma_0 & \Pi \end{bmatrix} \begin{bmatrix} s_{t-1} \\ \tilde{X}_{t-1} \end{bmatrix} + \begin{bmatrix} I & \mathbf{0} \\ \Gamma_0 & V \end{bmatrix} \begin{bmatrix} \epsilon_t \\ \xi_t \end{bmatrix}$$

The researcher obtains the posterior distribution for the joint time series y_t and then verifies for which components in B_{21} the 90% credible set doesn't include zero.

The relevance of the missing information from the extra observable variables can be analysed by a FEVD and by checking the proportion of the variance which is attributed to the missing observables.

3.4.2 Identification of misspecification by introducing uncorrelated exogenous processes

In Inoue, Kuo, and Rossi (2019), a method is proposed, consisting in adding uncorrelated exogenous processes, not without similarities to Chari, Kehoe, and McGrattan (2007). The fundamental difference is that in Chari, Kehoe, and McGrattan (2007), a very similar method, called Business Cycle Accounting approach, was used to improve the interpretability and the knowledge of the mechanisms in a DSGE model, which was not explicitly considered

as misspecified. The objective in Inoue, Kuo, and Rossi (2019) is to study the possibility of being in the presence of a misspecified model and of ways to correct it.

Let us assume the economist uses the following misspecified baseline model:

$$\begin{aligned} \underset{c_t}{\text{maximize}} \quad & E_0 \left[\sum_{t=0}^{\infty} \beta^t \left(\theta_0 c_t - \frac{\theta_1}{2} c_t^2 \right) \right] \\ \text{subject to} \quad & a_{t+1} = (1 + r_t)(a_t + y_t - c_t) \\ & y_t = y_{t-1} + \epsilon_t \end{aligned}$$

However, reality could have come from two other DGP, one in which the asset's return rate, r_t , is uncertain, and another in which the data on asset's value, a_t , has measurement error.

The idea is to add what the authors call *margin* variables, v_t and w_t to the baseline model:

$$\begin{aligned} \underset{c_t}{\text{maximize}} \quad & E_0 \left[\sum_{t=0}^{\infty} \beta^t \left(\theta_0 c_t - \frac{\theta_1}{2} c_t^2 \right) \right] \\ \text{subject to} \quad & a_{t+1} = (1 + r_t)(1 + v_{t+1})(a_t + w_t + y_t - c_t) \\ & y_t = y_{t-1} + \epsilon_t \\ & v_t = \eta_{v,t} \\ & w_t = \rho_w w_{t-1} + \eta_{w,t} \end{aligned}$$

where,

$$\begin{bmatrix} \epsilon_t \\ \eta_{v,t} \\ \eta_{w,t} \end{bmatrix} \stackrel{iid}{\sim} \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma_v^2 & 0 \\ 0 & 0 & \sigma_w^2 \end{bmatrix} \right)$$

In the world where r_t is uncertain, we are expecting v_t to be important in explaining the behaviour of the agent, and w_t to shut down. One can expect the reverse when the DGP is that of the measurement error. The researcher, thus, may add as many margin variables as desired, in thoughtful positions of his model.

Because the margin variables are exogenous processes, therefore not so dissimilar to shocks, one can use a FEVD analysis to ascertain the impact of the individual misspecifications.

Another technique to detect the source of misspecification is to analyse the marginal likelihood where the researcher eliminates one of the margins

at a time, and observe which margin elimination is responsible for the greatest decrease in the marginal likelihood, indicating a possible direction of improvement in the model.

One question poses itself — and to the best of our knowledge the literature has not yet considered — it is the context where if our baseline model is misspecified, could not some of the margins try to overcompensate for the lack of other margins, or the lack of other elements, pointing to misspecification in the wrong direction?

Except for BMA, so far, this is the first method which does not assume explicitly some kind of linearity. However, it is not clear what type of dynamics these margin variables can really capture, and the how solving of the DSGE will affect the margins.

3.4.3 Using A Composite Likelihood approach

In Canova and Matthes (2017), the authors propose to deal with model misspecification using a composite likelihood approach for estimating and doing inference in DSGE models. This method was originally developed when the likelihood of the full model was computationally intractable or difficult to construct.

Assuming we have an unknown DGP with density $f(z_t)$, and A_i a set of marginal or conditional events, with $f(z_{it} \in A_i, \theta, \eta_i)$ being the corresponding subdensities. These subdensities can be viewed as representing the likelihoods associated to different DSGE models in which we may be interested, or the densities generated by different approximate solutions to a model. There are several possible interpretations. Ultimately, the submodels considered may not even be statistically compatible.

For each A_i , we may have a different set of observables z_{it} , sample size T_i , and parameters $\psi_i = (\theta, \eta_i)$ where η_i are nuisance model specific parameters.

Given a vector of weights ω_i with $\sum \omega_i = 1$, which will determine the weight of each submodel, the composite likelihood is

$$\mathcal{CL}(\theta, \{\eta_i\}_{1,\dots,K}, \{z_{it}\}_{t=1,\dots,T_i}^{i=1,\dots,K}) = \prod_{i=1}^K [\mathcal{L}(\theta, \eta_i | z_{it} \in A_i)]^{\omega_i}$$

The above expression is not a likelihood, and seems to bear some resemblance to the BMA approach. Despite that, one of the main difference from the Composite Likelihood regarding BMA is that θ will be estimated using the information present in *all* of the submodels, whereas in BMA estimation is done for each submodel independently. The observables vectors z_{it} may share some components with different submodels.

The authors show how to do a Bayesian Analysis with their method, mainly due to the fact that standard frequentist asymptotic theory needs

some regularity conditions, which are often violated in practice, and because Bayesian methods may help when the sample size is small.

The article then proceeds to give some examples on how the composite likelihood method may improve small sample and population identification problems, estimation when dealing with singular models, or how to robustly estimate structural parameters in different models.

Similarly to the BMA, we could also ask ourselves what will happen when none of the models are well specified. In this case, no corrective measure for misspecification is given by the authors.

3.4.4 Agnostic Structural Disturbances

The method of adding agnostic structural disturbances (ASD), recently proposed in Haan and Drechsel (2018), focuses on misspecification originating *specifically* from structural disturbances, and it has two equivalent formulations, which we will see below. To implement this method one must consider a *linearised* model.

First, we will make clear the distinction between the notion of a measurement error and that of a structural disturbance. With this objective in mind, consider the following system:

$$\begin{aligned} s_t &= As_{t-1} + Bw_t \\ z_t &= Cs_{t-1} + Dw_t \\ w_t &= Gw_t + \eta_t \end{aligned}$$

To think of w_t as a measurement error, one would have $B = 0$ and $D \neq 0$, whereas for a disturbance error, we would have $B \neq 0$ and $D = 0$. Contrary to the measurement error, the structural error affects the economic variables, and is propagated through the dynamics, by the state equation. The idea of an agnostic procedure is to add different structural disturbances in such a way that it does not impose additional restrictions on policy rules (*agnostic*), which is what may happen when a simple structural disturbance is added to the model equations.

Adding Disturbances to Model Equations

A linearised model may be represented in the following general way:

$$\begin{aligned} 0 &= \mathbb{E}_t (\Lambda_2(\Psi)s_{t+1} + \Lambda_1(\Psi)s_t + \Lambda_0(\Psi)s_{t-1} + \Gamma(\Psi)\epsilon_{t+1} + \Phi(\Psi)\epsilon_t) \\ &= \mathbb{E}_t \left[\begin{array}{c} \Lambda_2(\Psi)s_{t+1} + \Lambda_1(\Psi)s_t + \Lambda_0(\Psi)s_{t-1} \\ + [\Gamma_1(\Psi) \quad \Gamma_2(\Psi)] \begin{bmatrix} \epsilon_{1,t+1} \\ \epsilon_{2,t+1} \end{bmatrix} + [\Phi_1(\Psi) \quad \Phi_2(\Psi)] \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{bmatrix} \end{array} \right] \end{aligned}$$

where Ψ is a vector containing all the structural parameters, s_t is the state vector, and ϵ_t representing the exogenous random variables.

Let us assume that the researcher was only sure of a subset m_1 of the disturbances in ϵ_t , $\epsilon_{1,t}$ of dimension $m_1 \times 1$, and with $\dim(s_t) = n \times 1$. If $m_1 < n$ and there are no other disturbances, then we are in the presence of a singularity problem. One usual procedure is to add measurement errors. Another possibility would be to add structural disturbances, which in this case has been represented by $\epsilon_{2,t}$ of dimension $m_2 \times 1$. The submatrices $\Gamma_2(\Psi)$ and $\Phi_2(\Psi)$ capture the restrictions imposed by the additional m_2 structural disturbances. For the structural disturbances to be agnostic, the authors define $\Gamma_2(\Psi)$ and $\Phi_2(\Psi)$ as being independent of Ψ .

Adding Disturbances to Policy Functions

The solution to the above model could also be written in the following representation

$$\begin{aligned} s_t &= A(\Psi)s_{t-1} + [B_1(\Psi) \quad B_2(\Psi)] \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{bmatrix} \\ &= \sum_{i=1}^m s_t^{[i]}, \\ s_t^{[i]} &= A(\Psi)s_{t-1}^{[i]} + B_{.,i}(\Psi)\epsilon_{i,t} \end{aligned}$$

where $s_t^{[i]}$ can be seen as the outcome as if the economy had only just disturbance $\epsilon_{i,t}$, and $B_{.,i}$ the i -th column of $B(\Psi)$. It is of a paramount importance the assumption of linearity in the model specification. Otherwise, the equivalence of these two representations is no longer warranted. A proposition from the paper states that in this formulation, when the additional structural $\epsilon_{2,t}$ is not correlated with $\epsilon_{1,t-1}$, then the policy A and submatrix B_1 are invariant to Γ_2 and Φ_2 , the submatrices which characterize the agnostic structural disturbances.

Adding $m_2 = m - m_1$ structural agnostic disturbances will result in

$$\begin{aligned} s_t &= \sum_{i=1}^m s_t^{[i]}, \\ s_t^{[i]} &= A(\Psi)s_{t-1}^{[i]} + B_{\cdot,i}(\Psi)\epsilon_{i,t} \text{ for } i \leq m_1, \\ s_t^{[i]} &= A(\Psi)s_{t-1}^{[i]} + \hat{B}_{\cdot,i}\epsilon_{i,t} \text{ for } m_1 + 1 \leq i \leq m, \end{aligned}$$

Similarly to the previous formulation, $\hat{B}_{\cdot,i}$ will also be independent of the structural parameters, being associated to an agnostic disturbance rather than a regular structural disturbance. As a consequence there will be no global identification of $\hat{B}_{\cdot,i}$ since with ASDs, there is no distinguishing between the different structural disturbances, i.e. no distinguishing between $\epsilon_{i,t}$ and $\epsilon_{j,t}$, $\forall i, j \in [m_1 + 1, m]$. This relieves the researcher from the worry of having to be very specific about the behaviour of the structural disturbances. When the wedges of Chari, Kehoe, and McGrattan (2007), or marginal variables of Inoue, Kuo, and Rossi (2019), are added, one must determine in which equation the disturbance appears, and in what way it will interact with the parameter vector Ψ , and usually are only allowed to appear on a subset of the equations. This is a shortcoming relative to the ASD method.

In Haan and Drechsel (2018) it is also shown how one can test for model misspecification, by using a test, such as a likelihood-ratio test, from the perspective of model selection, i.e. we compare a model, and that same model with one of the disturbances replaced by an ASD. Some DGP simulations presented in the paper showed that the ASD approach was able to perform well in correcting the estimation of Ψ .

By focusing on just disturbance misspecification, this method adds less parameters to be estimated, making it much more parsimonious than methods such as the Hybrid Models approach. However, that which grants its parsimony, it is also another source of its shortcomings. This procedure was not conceived to detect, or correct, sources of misspecification disconnected from structural disturbances.

Chapter 4

UQ of Model Discrepancy with a Full Gaussian Process Prior

4.1 The Landscape

Let us do a general summary of what we have analysed so far, and in this way it will be easier to contextualise this chapter.

In Chapter 1, we tried to give a short overall description of how the Macroeconomics subject has behaved, and how it reacted to some important events that occurred in the last 2 decades, namely the period of great moderation, and then the great financial crisis. One of the main conclusions was that many of the lead researchers in the area, if not most, had become very critical of the usual uses of DSGE models, pointing at several theoretical modelling assumptions as the culprits. This was a recognition of serious misspecification. Even if those assumptions are relaxed, and theory is furthered, we may never reach the state where models are tractable and 'almost' correctly specified to be of any meaningful use.

In Chapter 2, we saw how the formal techniques of UQ were starting to be applied to Macroeconomics, and no one had yet tried to account for model uncertainty regarding DSGE modelling, using the idea of Kennedy and O'Hagan (2001), that of adding a bias term with a GP prior.

In Chapter 3, several methods to deal with misspecification were presented critically, taking stock of their limitations, which typically are the assumption of linearity in the SSM representing the DSGE; the lack of a proposed corrective procedure when none of the models were correctly specified; or even the restricted dynamics they could capture and correct for.

In this chapter, we will focus on applying, from a black-box perspective, an adaptation of the method of Kennedy and O'Hagan (2001) to DSGE

models. This approach is applicable to non-linear SSM, and is able to capture very diverse dynamics. With the objective of comparing with some state of the art methods for correcting misspecification, we look at Ireland (2004). In this reference, Ireland adds a VAR error term to the observation equation of the linearized DSGE model. His purpose was two-fold: first to be able to solve the model without needing to add structural disturbances; second, correct for possible misspecification by estimating structural parameters more in line with previous scientific studies. Hence, inspired by all these references, our method is to add to the observation equation of the SSM a discrepancy term with a Gaussian Process prior to correct the misspecification of the DSGE, whilst hoping to obtain a better interpretability of the structural parameters than that of Ireland (2004). However, in some cases, even with somewhat simpler models, a discrepancy term such as this may still lack the power to overcome the gap between the economic model and reality, as Brynjarsdóttir and O'Hagan (2014) exemplify.

This chapter is organized as following: in section 4.2, we start by stating the model we will be using to account for model uncertainty, and how we will quantify the bias term, either into the future, or in the past, for non-observed data. This will raise the question on how we will learn the model. In section 4.3, we show how, using a Gibbs like algorithm, we can sample from the posterior distribution of θ conditioned on particle $x_{1:T}$, using a Random-Walk Metropolis-Hastings (RWMH) step, and then from the state-space using a Particle Gibbs Ancestor Sampling (PGAS) algorithm. At the end of this same section, we specify the economic model we will be analysing for model discrepancy, and the GP characterising the bias term. Finally, in section 4.4, we present some results, focused mainly on forecasting, since it is usually one of the most interesting exercises in macroeconomics, more akin to extrapolation, even though our method may also be used for interpolation instead. At the end of this chapter, in section 4.5, we will present some succinct conclusions, with the main purpose to take some stock, and motivate the work in Chapter 5.

4.2 Quantifying Structural Uncertainty in a DSGE Model

4.2.1 A Bias Term with a Gaussian Process Prior

In this subsection, we will present the model we will be using for quantifying model discrepancy. Although, in what follows, we will assume linearity in the SSM of a DSGE for ease of exposition, our method may be applied to

non-linear models as well, as long as we are able to specify the matrices in the SSM as time-invariant. As such, we may have the following state-space representation:

$$\begin{aligned}x_t &= Ax_{t-1} + B\epsilon_t \\ y_t &= C_* + Cx_t\end{aligned}$$

Again for ease of exposition, we hide the dependency of A , C_* and C of the structural parameters of the DSGE model, which throughout this chapter will be represented by θ . In later sections, when we estimate the models, we will pay greater attention to θ , but for now there is no need.

Ireland (2004), uses as a baseline model the Real Business Cycle (RBC) model in Hansen (1985). In this RBC model, the x_t is a vector of unobserved variables, where each component is the log deviation, from its steady-state level, of capital and the technology shock. The y_t vector has as its components the output, the numbers of hours worked and consumption, all also in log deviations from their steady-state.

The DSGE model presented in Hansen (1985), uses only one shock to drive all the fluctuations, while having y of a dimension greater than one, and as such, by B. F. Ingram, Kocherlakota, and Savin (1994), the model is stochastically singular, and any estimation by maximum likelihood fails. A common approach to make the model amenable is to add structural shocks until the number of shocks equals the number of observable variables used in estimation.¹ An alternative method to estimate it would be the one used in Ireland (2004).

$$\begin{aligned}x_t &= Ax_{t-1} + B\epsilon_t \\ y_t &= C_* + Cx_t + u_t \\ u_t &= Du_{t-1} + \xi_t\end{aligned}$$

where $\xi_t \stackrel{iid}{\sim} N(0, V)$. So, Ireland adds a VAR error term to the observation equation in order to be able to estimate the model. The VAR term can be interpreted as a measurement error term. Although as a method of solving the model, it is not the most commonly used, from a misspecification perspective and specially for quantifying uncertainty with a discrepancy term, his results are somewhat positive and promising, since he is able to obtain more reasonable estimates for the structural parameters, more in line with other empirical studies, than it would be possible without his VAR term.

The original model studied in Ireland (2004) has B as a not full row rank matrix and this causes a singularity issue, which would also impede

¹See also Ireland (2004) for examples using this method.

the use of our method. Because the purpose of our work is not to find new ways to estimate the model itself, or to create an improved economic model to explain the data, but to quantify the misspecification inherent to each model, we will simply overcome this difficulty by adding an extra shock to the original transition equation, somehow not completely foreign to the ASD approach which was presented in the previous chapter, in subsection 3.4.4. Hence,

$$\begin{aligned}x_t &= Ax_{t-1} + \eta_t \\ y_t &= C_* + Cx_t + b(x_t) + e_t\end{aligned}$$

where $\eta_t \stackrel{iid}{\sim} N(0, Q)$, and we have added a discrepancy term to the observation equation to measure model bias/misspecification, and with $b(x) \sim \mathcal{GP}(m(x), K(x, x'))$ and $e_t \sim N(0, \Sigma)$, where Q and Σ are diagonal.

4.2.2 Prediction in a DSGE Model with a GP Bias Term

In this section, we give a more detailed description of the stochastics driving the model with a bias term, which will be necessary for the accounting of model uncertainty. This description is inspired by the exposition in Frigola-Alcalde et al. (2013) and Frigola-Alcalde (2015). A reader unfamiliar with GP, may find more useful to first read our appendix on GPs, and only then dive on this exposition, since we will be using several notions related to GP regression from a multi-output perspective.

As a consequence of our modelling in the previous subsection, $x_1 \sim p(x_1)$, $x_{t+1}|x_t \sim N(x_{t+1}|Ax_t, Q)$ and $y_t|x_t, b_t \sim N(y_t|C_* + Cx_t + b_t, \Sigma)$, where $b_t = b(x_t)$

The full joint probability has the following form:

$$\begin{aligned}p(y_{1:T}, x_{1:T}, b_{1:T}) &= p(y_T|y_{1:T-1}, x_{1:T}, b_{1:T})p(y_{1:T-1}, x_{0:T}, b_{1:T}) \\ &= p(x_{1:T})p(b_{1:T}|x_{1:T}) \prod_{t=1}^T p(y_t|x_t, b_t) \\ &= p(x_1) \left(\prod_{t=2}^T p(x_t|x_{t-1}) \right) N(b_{1:T}|m(x_{1:T}), K(x_{1:T}, x_{1:T})) \\ &\quad \cdot \prod_{t=1}^T p(y_t|x_t, b_t)\end{aligned}$$

To estimate $b_* = b(x_*)$ for an arbitrary x_* , with θ being the parameters of

the model,

$$p(b_* | x_*, y_{1:T}) = \int p(b_* | x_*, x_{1:T}, y_{1:T}, \theta) p(x_{1:T}, \theta | x_*, y_{1:T}) dx_{1:T} d\theta$$

and doing a Monte Carlo approximation, we get

$$p(b_* | x_*, y_{1:T}) \approx \frac{1}{N} \sum_{i=1}^N p(b_* | x_*, x_{1:T}[i], y_{1:T}, \theta[i])$$

To find $p(b_* | x_*, x_{1:T}, y_{1:T}, \theta) = p(b_* | x_*, x_{1:T}, \theta)$, we notice that $y_t = \tilde{b}_t + e_t$, where $\tilde{b}_t \sim \mathcal{GP}(m(x_t) + C_* + Cx_t, K(x_t, x_t))$. Hence, using the formulas for a GP regression, we obtain $p(\tilde{b}_* | x_*, x_{1:T}, \theta) = N(\tilde{b}_* | \hat{\tilde{b}}_*, \hat{\Sigma}_*)$ with

$$\begin{aligned} \hat{\tilde{b}}_* &= m(x_*) + C_* + Cx_* \\ &+ K(x_*, x_{1:T})(K(x_{1:T}, x_{1:T}) + I_T \otimes \Sigma)^{-1} (y_{1:T} - m(x_{1:T}) - (I_T \otimes C)x_{1:T} - C_{*1:T}) \end{aligned}$$

and

$$\hat{\Sigma}_* = K(x_*, x_*) - K(x_*, x_{1:T})(K(x_{1:T}, x_{1:T}) + I_T \otimes \Sigma)^{-1} K(x_*, x_{1:T})^\top$$

where $C_{*1:T} = (\mathbf{1}_{T \times 1} \otimes I_D)C_*$. This is the predictive distribution of a Multi-Output Gaussian Process Regression model, with $x_{1:T}$ as input, and $y_{1:T}$ as noisy output.² Therefore $p(b_* | x_*, x_{1:T}, \theta) = N(b_* | \hat{\tilde{b}}_* - Cx_* - C_*, \hat{\Sigma}_*)$. Hence, the above approximation results in

$$p(b_* | x_*, y_{1:T}) \approx \frac{1}{N} \sum_{i=1}^N N(b_* | \hat{\tilde{b}}_*[i] - C[i]x_* - C_*[i], \hat{\Sigma}_*[i])$$

Doing a similar reasoning for y_* , we get

$$\begin{aligned} p(y_* | x_*, y_{1:T}) &= \int p(y_* | x_*, y_{1:T}, \tilde{b}_*, \theta) p(\tilde{b}_* | x_*, y_{1:T}, \theta, x_{1:T}) p(\theta, x_{1:T} | y_{1:T}) d\tilde{b}_* d\theta dx_{1:T} \\ &\approx \frac{1}{N_y} \sum_{i=1}^{N_y} p(y_* | \tilde{b}_*[i], \Sigma[i]) \\ &= \frac{1}{N_y} \sum_{i=1}^{N_y} N(y_* | \tilde{b}_*[i], \Sigma[i]) \end{aligned}$$

²See Appendix on Multi-Output Gaussian Process Regression. Although, in the notation of that appendix $y_{1:T} = \bar{y}$, we are abusing it regarding $K(x_{1:T}, x_{1:T})$, since here $x_{1:T} = \text{vec}(\tilde{X})$.

where for the sample $\{\tilde{b}_*[i], \Sigma[i]\}_{i=1:N_y}$, we will use the approximation given by $p(\tilde{b}_* | x_*, y_*) \approx 1/N \sum_{i=1}^N N(\tilde{b}_* | \hat{b}_*[i], \hat{\Sigma}[i])$. The number of mixture components N_y can be fixed independently of N (sample size of MCMC).

So far, we have not assumed explicitly whether we were doing a forecasting or not, even though in the experiments which we will be doing in section 4.4, the focus will be on forecasting. Instead of finding a specific x_* , one could marginalise over x_* . However, due to the computational burden that we will face, when doing simulations, we decided to use a specific value for x_* . More details can be found in section 4.4.

To sample from $p(x_{1:T}, \theta | x_*, y_{1:T})$, we shall resort to an algorithm known as Particle Gibbs with Ancestor Sampling (PGAS), which was presented in Lindsten, Jordan, and Schön (2014). The PGAS is an algorithm specially well suited for addressing non-parametric Bayesian inference problems in State-Space models. More information about this algorithm and a small intuitive introduction to particle filtering can be found in the appendix on Bayesian Statistical Methods, at the end of our work.

4.3 Learning the Model

From the exposition in Lindsten, Jordan, and Schön (2014), we have **Algorithm 1** below.

Algorithm 1 PGAS for Bayesian Learning of SSMs

- 1: Set $\theta[0]$ and $x_{1:T}[0]$ from some distribution, and N as MCMC sample size
 - 2: **for** $n \in [1, N]$ **do**
 - 3: Draw $\theta[n] \sim p(\theta | x_{1:T}[n-1], y_{1:T}, \theta[n-1])$ /* Using Algorithm 2
 - 4: Draw $x_{1:T}[n] \sim p^N(x_{1:T} | x_{1:T}[n-1], y_{1:T}, \theta[n])$ /* Using Algorithm 3
 */
 - 5: **end for**
-

4.3.1 Sampling θ from its conditional posterior distribution

We can sample θ from the distribution $p(\theta | x_{1:T}[n], y_{1:T})$, by using a Metropolis-Hastings (MH) step, resulting in a Metropolis-within-Gibbs procedure. Starting from

$$p(\theta | x_{1:T}, y_{1:T}) \propto p(x_{1:T}, y_{1:T} | \theta) p(\theta),$$

and then marginalizing over \tilde{b} GP, and using the formula for the full joint probability developed in section 4.2.2, we get

$$\begin{aligned}
p(x_{1:T}, y_{1:T} | \theta) p(\theta) &= p(\theta) \int p(x_{1:T}, y_{1:T}, \tilde{b}_{1:T} | \theta) d\tilde{b}_{1:T} \\
&= p(\theta) p_\theta(x_{1:T}) \int N(\tilde{b}_{1:T} | \tilde{m}(x_{1:T}), K(x_{1:T}, x_{1:T})) \prod_{t=1}^T N(y_t | \tilde{b}_t, \Sigma) d\tilde{b}_{1:T} \\
&= p(\theta) p_\theta(x_{1:T}) \int N(\tilde{b}_{1:T} | \tilde{m}(x_{1:T}), K(x_{1:T}, x_{1:T})) N(y_{1:T} | \tilde{b}_{1:T}, I_T \otimes \Sigma) d\tilde{b}_{1:T} \\
&= p(\theta) p_\theta(x_{1:T}) |(2\pi)^{DT} \tilde{K}_{1:T}|^{-1/2} e^{(-\frac{1}{2}(y_{1:T} - \tilde{m}(x_{1:T}))^\top \tilde{K}_{1:T}^{-1} (y_{1:T} - \tilde{m}(x_{1:T})))} \\
&= p(\theta) p_\theta(x_{1:T}) N(y_{1:T} | \tilde{m}(x_{1:T}), \tilde{K}_{1:T})
\end{aligned}$$

where

$$\tilde{m}(x_{1:T}) = m(x_{1:T}) + (I_T \otimes C)x_{1:T} + \begin{bmatrix} I_D \\ \vdots \\ I_D \end{bmatrix} C_*$$

and

$$\tilde{K}_{1:T} = K(x_{1:T}, x_{1:T}) + I_T \otimes \Sigma$$

Now, using the equalities

$$p_\theta(x_{1:T}) = p_\theta(x_1) \prod_{t=2}^T N(x_t | Ax_{t-1}, Q) = p_\theta(x_1) N(x_{2:T} | (I_{T-1} \otimes A)x_{1:T-1}, I_{T-1} \otimes Q)$$

The marginal likelihood is equal to

$$\begin{aligned}
p(x_{1:T}[n], y_{1:T} | \theta) &= p_\theta(x_1) N(x_{2:T}[n] | (I_{T-1} \otimes A)x_{1:T-1}[n], I_{T-1} \otimes Q) \\
&\quad N(y_{1:T} | \tilde{m}(x_{1:T}[n]), \tilde{K}_{1:T})
\end{aligned}$$

Using the full likelihood derived above, we can compute the acceptance probability of the MH step in Algorithm 2, below. In this work, we will be using a modification of this proposal, and *blocking* the updates. For more information, check also subsection 4.3.4. We will use a MH step for each block $i \in \{1, \dots, B\}$, but building on the previous iteration, i.e. we will consider each time the following vector: $(\theta_{<i}[n], \theta_i^*, \theta_{>i}[n-1])$, with θ_i^* being the new draw for block i , $\theta_{<i}[n]$ the n -th draw for all blocks up to the i -th one, and $\theta_{>i}[n-1]$ being the previous iteration accepted draws for the remaining blocks.

Algorithm 2 Metropolis-Hastings Step

Input: $x_{1:T}[n]$, $y_{1:T}$, $\theta[n-1]$ **Output:** $\theta[n]$

- 1: **for** block $i \in \{1, \dots, B\}$ **do**
- 2: Draw θ_i^* from their respective proposal density
 $q_i(\theta_i^* | \theta_{<i}[n], \theta_i^*, \theta_{>i}[n-1], x, y)$
- 3: Define $A = \min(1, f)$ with

$$f = \frac{p(x, y | \theta_{<i}[n], \theta_i^*, \theta_{>i}[n-1]) \cdot p(\theta_{<i}[n], \theta_i^*, \theta_{>i}[n-1])}{p(x, y | \theta_{<i}[n], \theta_{\geq i}[n-1]) \cdot p(\theta_{<i}[n], \theta_{\geq i}[n-1])} \cdot \frac{q(\theta_i[n-1] | \theta_{<i}[n], \theta_i^*, \theta_{>i}[n-1], x, y)}{q(\theta_i^* | \theta_{<i}[n], \theta_{\geq i}[n-1], x, y)}$$

- 4: Draw $u_i \sim \text{Unif}[0, 1]$
 - 5: **if** $u_i \leq A$ **then**
 - 6: $\theta_i[n] = \theta_i^*$
 - 7: **else**
 - 8: $\theta_i[n] = \theta_i[n-1]$
 - 9: **end if**
 - 10: **end for**
-

4.3.2 Sampling from the State-Space

Adapting the PGAS algorithm to our setting, following Lindsten, Jordan, and Schön (2014) and Frigola-Alcalde (2015), we obtain Algorithm 3. Before presenting the algorithm, some explanations are in order. First, the Particle Gibbs(PG) sampler is based on a Sequential Monte Carlo(SMC) in which a reference trajectory is fixed at onset. A proposal is used to sample x_t^i from a distribution, then each particle is assigned a weight, and we resample from those particles according to these weights. The resampling is done by choosing the ancestor particle $x_{1:t-1}^{a_t^i}$, and then defining the new particle as $x_{1:t}^i = (x_{1:t-1}^{a_t^i}, \tilde{x}_t^i)$. However, the PG has a serious limitation, since we may have path degeneracy in the SMC, resulting in a poor mixing of the Markov Kernel. The problem of path degeneracy is common and is more likely with increasing dimensionality of the problem at hand. The usual tweak to this has been to introduce a backward simulation step in the PG. However, this additional step imposes some restrictions on the allowed dependencies of the model being studied. To allow for a non-parametric Bayesian analysis, the PGAS tries to improve on this issue by modifying the PG with one extra

step where we sample the ancestor of the *reference* particle³.

Algorithm 3 PGAS Markov Kernel

Input: $x_{1:T}[n-1]$, $\theta[n-1]$, and N_p which is the number of particles

Output: $x_{1:T}[n]$

- 1: Set $\tilde{x}_{1:T} = x_{1:T}[n-1]$ as the reference trajectory
- 2: Draw $x_1^i \sim p(x_1|\theta[n-1])$ for $i = 1, \dots, N_p - 1$.
- 3: Set $x_1^N = \tilde{x}_1$
- 4: Set $w_1^i = \frac{1}{N_p}$ for $i = 1, \dots, N_p - 1$.
- 5: **for** $t = 2, \dots, T$ **do**
- 6: Draw a_t^i with $P(a_t^i = j) \propto w_{t-1}^j$ for $i = 1, \dots, N_p - 1$.
- 7: Draw $x_t^i \sim p(x_t|x_{1:t-1}^{a_t^i}, \theta[n-1])$ for $i = 1, \dots, N_p - 1$.
- 8: Compute, with parallelization, $\{\tilde{w}_{t-1|T}^i\}_{i=1}^{N_p}$ using

$$\tilde{w}_{t-1|T}^i = w_{t-1}^i \frac{p_{\theta[n-1]}((x_{1:t-1}^i, \tilde{x}_{t:T}), y_{1:T})}{p_{\theta[n-1]}(x_{1:t-1}^i, y_{1:t-1})}$$

- 9: Draw $a_t^{N_p}$ with $P(a_t^{N_p} = j) \propto \tilde{w}_{t-1|T}^j$
 - 10: Set $x_t^{N_p} = \tilde{x}_t$
 - 11: Set $x_{1:t}^i = (x_{1:t-1}^{a_t^i}, x_t^i)$ for $i = 1, \dots, N_p$.
 - 12: Compute, using parallelization, $w_t^i = \frac{p_{\theta[n-1]}(x_{1:t}^i, y_{1:t})}{p_{\theta[n-1]}(x_{1:t-1}^i, y_{1:t-1}) \cdot r_{\theta[n-1]}(x_t^i|x_{1:t-1}^i)}$
for $i = 1, \dots, N_p$.
 - 13: **end for**
 - 14: Sample k with $P(k = i) \propto w_T^i$ and set $x_{1:T}[n] = x_{1:T}^k$
-

A few words are in order to clarify and simplify the algorithm, namely in steps 7, 8 and 12. In Step 7, it is useful to notice that contrary to Frigola-Alcalde (2015) we don't have a GP in the transition equation, only in the observation equation, hence the density simplifies considerably to

$$p(x_t|x_{1:t-1}^{a_t^i}, \theta[l-1]) = N(A[l-1]x_{t-1}^{a_t^i}, Q[l-1]).$$

In step 8, the formula for the sequential computing of the weight for the particles can be simplified a bit, by noticing that

³In Algorithm 3, this extra step is n^o9. More information can be found on Appendix B, on Bayesian Statistical Methods

$$\begin{aligned}\tilde{w}_{t-1|T}^i &= w_{t-1}^i \frac{p_{\theta[n-1]}((x_{1:t-1}^i, \tilde{x}_{t:T}), y_{1:T})}{p_{\theta[n-1]}(x_{1:t-1}^i, y_{1:t-1})} \\ &\propto w_{t-1}^i f_{\theta}(\tilde{x}_t | x_{t-1}^i) \frac{N(y_{1:T} | \tilde{m}(x_{1:t-1}^i, \tilde{x}_{t:T}), \tilde{K}(x_{1:t-1}^i, \tilde{x}_{t:T}))}{N(y_{1:t-1} | \tilde{m}(x_{1:t-1}^i), \tilde{K}(x_{1:t-1}^i))}\end{aligned}$$

where the proportional symbol is used to indicate a missing factor which does not depend on the index i , i.e. it's a constant along all the particles for a given iteration in Algorithm 3.

In Step 12, we have $r_{\theta}(x_t | x_{t-1})$ which is a chosen proposal density. If for this proposal we use the transition density, then we have

$$\frac{p_{\theta[n-1]}(x_{1:t}^i, y_{1:t})}{p_{\theta[n-1]}(x_{1:t-1}^i, y_{1:t-1}) \cdot r_{\theta[n-1]}(x_t^i | x_{1:t-1}^i)} = \frac{N(y_{1:t} | \tilde{m}(x_{1:t}^i), \tilde{K}(x_{1:t}^i))}{N(y_{1:t-1} | \tilde{m}(x_{1:t-1}^i), \tilde{K}(x_{1:t-1}^i))}.$$

In the two previous fractions with normal densities, one can alleviate their computational burden slightly by noticing that $p(y_2 | y_1) = \frac{p(y_1, y_2)}{p(y_1)}$, and for

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N \left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right) \text{ we have}$$

$$x_2 | x_1 \sim N \left(\mu_2 + \Sigma_{21} \Sigma_{11}^{-1} (x_1 - m_1), \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \right).$$

The original algorithm does not use parallelisation of commands, but we see a great increase in speed of computations when we do so. In this case, we refer to something more akin to *data* parallelism, i.e. contemporaneous/simultaneous computation of the same task on different components of data. Here, we are computing the weights, using different data (particles' history). This is different from another sense of parallelisation, where we run different tasks simultaneously across multiple processors, with the same, or different, data, i.e. *task* parallelisation. See for example Quinn (2003). Similarly to any other computer programme, its computational time will depend on the specificities of the computer being used. The more cores the mainframe/server has, the faster the programme will run, when we use parallelization.

4.3.3 The DSGE Model

The DSGE Model studied in Ireland (2004) is a prototypical DSGE model of that time, i.e., an RBC model, where nominal variables have no influence over real quantities. In this subsection, two small abuses of notation will be

made. Here, the parameter θ will not comprehend all the parameters, but as in Ireland (2004), it will refer to the capital's share of output. The remaining parameters of the economic model are $\beta, \delta, \rho, \eta, \gamma$, and A . Likewise, y_t , for the purposes of this subsection, will signify *output*, in the economic sense. Given these warnings, we continue with the exposition of the RBC model. The SSM may be solved to have the following form:

$$\begin{aligned}x_{t+1} &= \Pi x_t + W \epsilon_{t+1} \\ \tilde{z}_{t+1} &= U x_t\end{aligned}$$

where $x_t = (\hat{k}_t \ \hat{a}_t)^\top$, $\tilde{z}_t = (\hat{y}_t \ \hat{\iota}_t \ \hat{h}_t \ \hat{c}_t)^\top$, with the hat notation meaning log deviations from steady-state. The x_t is a vector of unobserved variables, where each component is the log deviation, from its steady-state level, of capital (k_t) and the technology shock (a_t). The \tilde{z}_t vector has as its components the output (y_t), investment (ι), the numbers of hours worked (h_t) and consumption (c_t), all also in log deviations from their steady-state. The matrices from the SSM can be described using the following formulas $\Pi = \begin{bmatrix} S_3 & S_4 \\ 0 & \rho \end{bmatrix}$,

$$U = \begin{bmatrix} S_5 & S_6 \\ S_1 & S_2 \end{bmatrix}, \quad W = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

$$\begin{aligned}
S_1 &= \frac{K_{22} - K_{11}}{K_{12}} \\
S_2 &= \frac{(K_{22} - K_{11})L_1 - K_{12}L_2}{K_{12}(K_{11} - \rho)} \\
S_3 &= K_{22} \\
S_4 &= \frac{K_{12}L_2}{K_{22} - K_{11}} + \frac{(K_{22} - \rho)((K_{22} - K_{11})L_1 - K_{12}L_2)}{(K_{22} - K_{11})(K_{11} - \rho)} \\
S_5 &= \begin{bmatrix} 1 - \frac{1-\theta}{\theta}S_1 \\ S_1 + \left(\frac{\eta/\beta-1+\delta}{\theta^2(\eta-1+\delta)}\right)(\theta - S_1) \\ 1 - \frac{1}{\theta}S_1 \end{bmatrix} \\
S_6 &= \begin{bmatrix} \frac{1}{\theta} - \frac{1-\theta}{\theta}S_2 \\ S_2 + \left(\frac{\eta/\beta-1+\delta}{\theta^2(\eta-1+\delta)}\right)(1 - S_2) \\ \frac{1}{\theta} - \frac{1}{\theta}S_2 \end{bmatrix} \\
K_{11} &= \frac{\eta - \beta(1 - \theta)(1 - \delta)}{\beta\eta\theta} \\
K_{12} &= \frac{\beta\eta\theta^2 - \eta + \beta(1 - \theta^2)(1 - \delta)}{\beta\eta\theta^2} \\
K_{22} &= \frac{\eta\theta}{\eta - \beta(1 - \theta)(1 - \delta)} \\
L_1 &= \frac{\eta - \beta(1 - \delta)}{\beta\eta\theta^2} \\
L_2 &= \frac{\rho(\eta - \beta(1 - \delta))}{\eta - \beta(1 - \theta)(1 - \delta)}
\end{aligned}$$

However, since by construction of the data we always have $Y_t = C_t + I_t$ (output at each period in a closed economy is defined as consumption plus investment), our model can be rewritten in fact as

$$\begin{aligned}
x_{t+1} &= \Pi x_t + W \epsilon_{t+1} \\
\hat{z}_{t+1} &= C x_t
\end{aligned}$$

where $\hat{z}_t = (\hat{y}_t, \hat{h}_t, \hat{c}_t)$ (the log-deviations from the steady state for each respective variable) and C is the matrix obtained by deleting the 2nd row of the previous U matrix.

The observed data, taken from the FRED database, maps output Y_t , consumption C_t and H_t hours worked to some respective time-series. The

RBC implies that all of these three variables will be growing at a constant rate in the steady state⁴, so we must detrend the data. Hence, we have $\hat{y}_t = \log(Y_t) - t \log(\eta) - \log(y_*)$ where y_* is the steady state of the RBC model for $\frac{Y_t}{\eta^t}$. Similarly, we have $\hat{c}_t = \log(C_t) - t \log(\eta) - \log(c_*)$ and $\hat{h}_t = \log(H_t) - \log(h_*)$. The steady state values for each of the above can be restated as a function of the structural economic parameters:

$$\begin{aligned} c_* &= \left(1 - \frac{\theta(\eta - 1 + \delta)}{\eta/\beta - 1 + \delta}\right) y_* \\ h_* &= \frac{1 - \theta}{\gamma} \left(1 - \frac{\theta(\eta - 1 + \delta)}{\eta/\beta - 1 + \delta}\right)^{-1} \\ y_* &= A^{1/(1-\theta)} \left(\frac{\theta}{\eta/\beta - 1 + \delta}\right)^{\theta/(1-\theta)} \frac{1 - \theta}{\gamma} \left(1 - \frac{\theta(\eta - 1 + \delta)}{\eta/\beta - 1 + \delta}\right)^{-1} \end{aligned}$$

This results in the rewriting of the SS representation for the RBC model as

$$\begin{aligned} x_{t+1} &= \Pi x_t + W \epsilon_{t+1} \\ \hat{m}_{t+1} &= C_*(t) + C x_t \end{aligned}$$

with $\hat{m}_t = \begin{bmatrix} \log(Y_t) \\ \log(H_t) \\ \log(C_t) \end{bmatrix}$ being the logarithm of the observed time-series, and $C_*(t) = t \begin{bmatrix} \log(\eta) \\ 0 \\ \log(\eta) \end{bmatrix} + \begin{bmatrix} \log(y_*) \\ \log(h_*) \\ \log(c_*) \end{bmatrix}$. We can notice that the SSM has become non-autonomous. Our procedure will be to push the time variant component to the left hand side of the equation, allowing the system to be time invariant. Also, similarly to the common practice of calibration in Macroeconomics⁵, we shall fix η at the value which was estimated by Ireland (2004), i.e. $\eta = 1.0051$ ⁶.

Rewriting the SSM representation for the RBC model as

$$\begin{aligned} x_{t+1} &= \Pi x_t + W \epsilon_{t+1} \\ \hat{m}_{t+1} &= C_* + C x_t \end{aligned}$$

⁴This is what means to be a in Balanced Growth Path

⁵In Ireland (2004), we have yet another example of this notion of calibration.

⁶In section 4.4.2 of this chapter, we will relax this assumption, by detrending the data using an updated value for η .

with

$$\hat{m}_t = \begin{bmatrix} \log(Y_t) - t \log(\eta) \\ \log(H_t) \\ \log(C_t) - t \log(\eta) \end{bmatrix},$$

and

$$C_* = \begin{bmatrix} \log(y_*) \\ \log(h_*) \\ \log(c_*) \end{bmatrix}.$$

In the following sections, \hat{m}_t will be referred instead by y_t , as the observable data, not to be confused with the output component.

Now, let us look at the economic interpretation of the structural parameters, which may guide our choice of priors. In Table 4.1, one can also see all the structural parameters, and their possible values.

Parameter	Economic Interpretation	Previous Studies' Values	Range
β	Discount Factor	0.95 upwards	$]0, 1[$
δ	Capital's Depreciation Rate	0.025 up to 0.05	$]0, 1[$
η	Growth Rate of Labour Augmenting Technology	1.004 up to 1.006	$]1, +\infty[$
θ	Capital's share in Output	0.30 up to 0.38	$]0, 1[$
ρ	Technology Shock's Persistence	0.99 upwards	$] - 1, 1[$
A	Steady-State for Technology Shock	Above 5	$]0, +\infty[$
γ	Proportion of Disutility of Hours Worked	Close to 0	$]0, +\infty[$

Table 4.1: A simple eliciting of priors for the structural parameters

In Table 4.1, the range column shows the theoretical allowed values for each parameter, and the previous studies' values shows prior information based on previous economic studies, such as Ireland (2004), Smets and Wouters (2003), and Smets and Wouters (2007), etc.⁷, that may help us elicit a prior distribution for the parameters. For the discount factor, what one usually expects is a value close to 0.95 with a preference for values closer to

⁷It is beyond the scope of this work to refer to the many scientific articles and the accompanying issues(time-varying, etc.) in the measurement of these values, for example in θ (income inequality). Suffice to say that, for the purposes of our work, we will use some standard textbook values.

1. For the USA, a usual expected value of θ is close to 0.35 with a preference for increasing values up to 0.4. The labour-augmenting technological progress for the US is considered to have a value close to 1.0045 to 1.006 giving a annual growth rate of 1.8% to 2.4%. The ρ parameter for RBC type of models is usually expected to reach values very close to 1, ≈ 0.99 , with a clear preference for upward values. The depreciation rate has values close to 0.025 with preference for values up to 0.05.

Plots for All Data

After applying the above transformations to the data used in Ireland (2004), we obtain the following plots

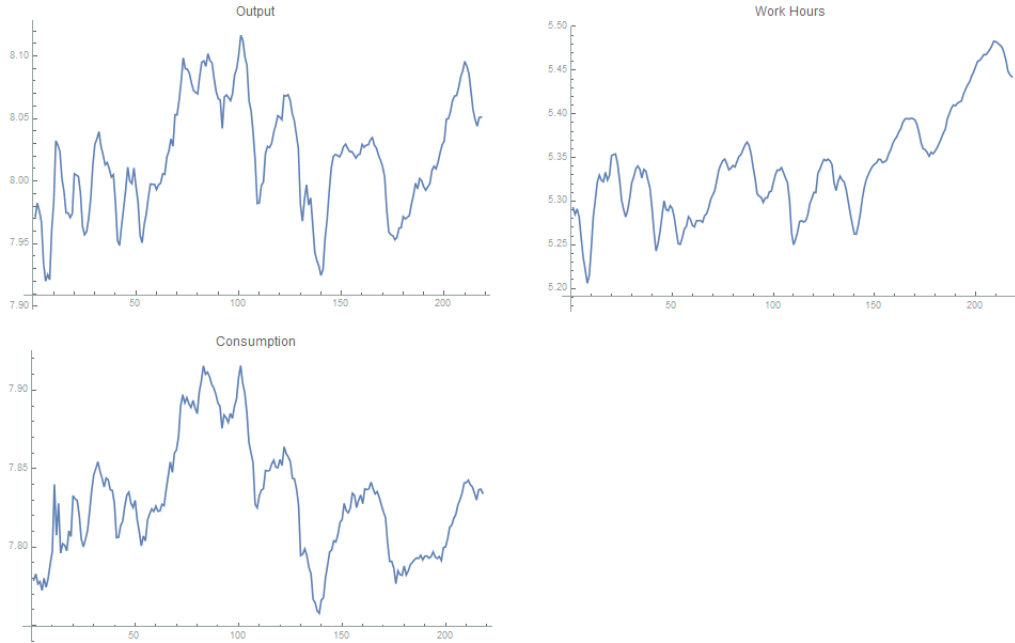


Figure 4.1: Plots for different components of \hat{m}

Observing Figure 4.1, one might be tempted to use it to help us guide our choice of Covariance Function, which determines the behaviour of the GP for the bias term. The plots seem very rough, comparable to what one could see when using a Matèrn covariance function with a low η . However, since the bias term is unobservable, one cannot be sure whether a rougher or smoother behaviour will be better. In Chapter 6, we will analyse this topic in greater detail.

4.3.4 Covariance Function, Priors and Proposal

Covariance Function

The Covariance Function to be chosen is of the Matérn Class. This type of covariance function allows for much rougher processes than the squared exponential, which makes the unidimensional Gaussian Process to be infinitely differentiable.⁸ The Matérn Class of functions is given by

$$k_M(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{r\sqrt{2\nu}}{l} \right)^\nu K_\nu \left(\frac{r\sqrt{2\nu}}{l} \right),$$

where $r = \|x - x'\|$. The function K_ν is the Modified Bessel of the 2nd kind. When $r = 0$, the right limit of $k_M(r)$ is 1. For fixed r and ν , k_M is increasing in l , and so l is known as the range or length-scale parameter. If we want farther observations to be more correlated, we use a greater value of l . Both ν and l are positive reals. For this class of functions the Gaussian Process is k -times differentiable if and only if $\nu > k$. When $\nu \rightarrow \infty$, we obtain the squared exponential function. For this reason ν is known as the smoothness parameter. We will start by allowing rougher processes, so as to allow for an increased richness in the function space. In this spirit, we fix $\nu = 0.5$, as is usual in the literature (see Rasmussen and Williams (2006)) for very rough processes, for which the Matérn function form simplifies to

$$k_M(r) = \exp^{-r/l}.$$

We will estimate from the data the value for the length-scale parameter.

One simple way to generalize the unidimensional Matérn Covariance function to a multidimensional setting is to notice that we may define a separable covariance function such as

$$\text{Cov}(f_d(x), f_{d'}(x')) = k(x, x')_{d,d'} = k_M(\|x - x'\|) \cdot k_D(d, d'),$$

which gives

$$\text{Cov}(f(x), f(x')) = k(x, x') = k_M(\|x - x'\|)M,$$

with $M_{ij} = k_D(i, j)$. Thus resulting in

$$K(X, X) = \sum_{l=1}^L k_M(X) \otimes M_l$$

⁸Differentiability, and continuity are here used in the Mean-Squared sense. See the appendix on GP, and also Rasmussen and Williams (2006)

where $k_M(X) = [k_M(\|x_i - x_j\|)]_{i,j \in \{1, \dots, T\}}$, and $\{M_l \in \mathbb{R}^{D \times D}\}$ are positive definite symmetric matrices.⁹ This is the same type of covariance function one would obtain if we were dealing with a linear model of coreginalization, and with $L = 1$ one obtains the intrinsic model of coregionalization¹⁰. In fact, due to some computational limitations, our analysis will be restrained to the case $L = 1$.

Priors

For the priors, there are at least two paths one could take, depending on how we deal with the likelihood and θ . The first takes advantage of $p(\theta|y, x) \propto p(\theta_x|x)P(\theta_y|y, x)$, where $\theta_x = (A, Q)$ and $\theta_y = (C, \theta_{gp}, \Sigma)$. For $p(x_{1:T}[n]|\theta_x)p(\theta_x)$, we notice that $x_t = Ax_{t-1} + \eta_t$ allows us to use the usual conjugate prior for the multivariate bayesian regression. For

$$p(\theta_y)p(y_{1:T}|x_{1:T}[n], \theta_y) = p(\theta_y)N(y_{1:T}|\tilde{m}(x_{1:T}[n]), \tilde{K}_{1:T})$$

we would use a MH step. We should notice that in this setting we are putting the priors $p(\theta_x)$ on the matrices of the transition equation — namely on A, B , etc. — instead of on the parameters with direct economic interpretation — namely ρ, γ , etc.

However, since the economic research praxis is putting priors directly on the economic parameters, we will consider jointly the θ_x and θ_y , with

$$p(\theta|x_{0:T}[n], y_{1:T}) \propto p(x_{0:T}[n], y_{1:T}|\theta)p(\theta)$$

and use a MH step with the full likelihood. Given the previous discussion on the parameter domains and their preferred values, we decided to use relatively vague priors to allow for some uncertainty on our GP modelling.

In Table 4.2, the parameter l is the length-scale parameter for the Matérn covariance function type, and we chose a distribution on the positive reals with a large sized standard deviation since there are no previous studies on which we could guide our algorithm, and so we decided for a somewhat diffuse prior. For the matrix M_1 in the covariance function, we decided to use Inv-Wish($I_3, 15$), which, after a few initial experiments, seemed to offer the best mixing for the algorithm.

⁹In some literature, since they define $f_{1:D}$ instead of $f_{1:T}$, we find $K(X, X) = \sum_{l=1}^L M_l \otimes k_M(X)$

¹⁰See for example H. Liu, J. Cai, and Ong (2018)

Parameter	Priors	Mean	Std Deviation
β	Beta(9.9, 1.7)	0.85	0.099
δ	Beta(1, 10)	0.091	0.083
θ	Beta(5, 10)	0.333	0.117
ρ	Beta(9.9, 1.7)	0.85	0.01
A	G(10, 0.5)	5	1.581
γ	G(0.25, 0.1)	0.025	0.05
l	G(2, 2)	4	2.828
σ_i^2	Inv-G(2.5, 0.5)	0.333	0.471
q_i^2	Inv-G(2.5, 0.5)	0.333	0.471

Table 4.2: Priors for the structural parameters

Proposal Distribution

Regarding the proposal for the MH step, it is common to use, at least in DSGE literature,

$$q(\theta[l]|\theta[l-1], y, x) = p_{t\text{-Student}}(\theta[l]|\mu(\theta[l-1]), \Sigma(\theta[l-1]), \delta)$$

where $p_{t\text{-Student}}(\cdot|\mu, \Sigma, \delta)$ is the density of Student-t distribution, and δ is the degrees of freedom which we will assume to be 3 for the remainder of our work, and the mean and variance follow

$$\begin{aligned}\mu(\theta) &= \theta + s_1 \Sigma(\theta) \frac{\partial}{\partial \theta} \text{Log}(p(y, x|\theta)) \\ \Sigma(\theta) &= s_2 V \\ V &= -\frac{\partial^2}{\partial \theta \partial \theta^\top} \text{Log}(p(y, x|\theta))^{-1}\end{aligned}$$

However, since we will be dealing with multidimensional input and output, some initial trials have shown it is too much of a computational burden to deal with the unevaluated expressions for the mean and the covariance in our model specially due to the presence of a determinant and an inversion in the proposal's density together with the Covariance function. Initial experiments, with $V = I$, and with recourse to matrix calculus for an explicit expression¹¹, and faster evaluation of the formulas above, we observed that due to numerical instabilities, the computation of $\mu(\theta)$ could not be ensured, putting at risk the convergence theory for the MCMC method. Therefore, we decided for a Random-Walk type of proposal in MH step.

¹¹See Appendix D

Let us define $\theta = (\theta_{\text{econ}}, \theta_{\text{var}}, M_1, l_{GP})$, where θ_{econ} is the vector with the 6 parameters with direct economic interpretation, θ_{var} is the vector with the main diagonal elements of the error covariance matrices, and l_{GP} is the range parameter of the Covariance function of the GP. The proposals for θ_{econ} , θ_{var} and l_{GP} will be a multivariate and a univariate t distribution:

$$q(\theta'_i | \theta_i, y, x) = p_T(\theta'_i | \mu(\theta_i), s_2 I_{\dim(\theta_i)})$$

$$\mu(\theta_i) = \theta_i + s_1 U$$

where $U \sim \text{Unif}(-1, 1)$, and s_1, s_2 are values to be defined by the user, in order to ensure a proper convergence to the posterior distribution.

Meanwhile, for M_1 we will impose an Inverse-Wishart distribution.¹²

$$q(M'_1 | \theta, y, x) = p_{IW}(M'_1 | \mu(M_1) \cdot (\nu + p + 1), \nu)$$

$$\mu(M_1) = M_1 + s_1 M$$

where $M \sim \text{Inv-Wish}(I_3, \nu)$ and $\nu = 15$. Contrary to the previous parameters, the update formula for the proposal does not depend on s_2 . Instead, that role is taken up by the ν . After some simulations, we found the value $\nu = 15$, and the use of the Mode, instead of the Mean, ensured a reasonable mixing.

Given this proposal, there is a mismatch between the parameters' support distribution and the proposal's support distribution, since one is a strict subset of the other. To improve the efficiency of the MH step regarding the acceptance rate, it is usually suggested to reparametrize the model in order to be sure the proposal draws do not consistently belong to an area of the parameter space where the prior assigns zero probability.

Initial simulations, without reparametrizing the model, showed a deficient acceptance rate of the proposal. Therefore, as a technique to improve the acceptance rate of the proposal to acceptable levels of 20% to 40%, we shall use the following reparametrization:

$$(\tilde{\beta}, \tilde{\delta}, \tilde{\phi}, \tilde{\rho}, \tilde{A}, \tilde{\gamma}, \tilde{l}, (\tilde{q}_i^2), (\tilde{\sigma}_i^2)) = F(\beta, \delta, \eta, \phi, \rho, A, \gamma, l, (q_i), (\sigma_i^2))$$

$$= (\text{logit}(\beta), \text{logit}(\delta), \text{logit}(\phi), \text{logit}(\rho), \log(A), \log(\gamma), \log(l), (\log(q_i^2)), (\log(\sigma_i^2)))$$

Even though ρ is in $] -1, 1[$, due to the prior taking only positive values in $]0, 1[$, the accepted values will necessarily be in $]0, 1[$.

This will result in the Jacobian

$$\left| \frac{\partial F^{-1}}{\partial \tilde{\theta}} \right| = \frac{e^{\tilde{\beta}}}{(1 + e^{\tilde{\beta}})^2} \frac{e^{\tilde{\delta}}}{(1 + e^{\tilde{\delta}})^2} \frac{e^{\tilde{\phi}}}{(1 + e^{\tilde{\phi}})^2} \frac{e^{\tilde{\rho}}}{(1 + e^{\tilde{\rho}})^2} \exp(\tilde{A} + \tilde{\gamma} + \tilde{l}) \prod_i^2 e^{\tilde{q}_i^2} \prod_j^3 e^{\tilde{\sigma}_j^2}$$

¹²If $X_{p \times p} \sim \text{Inv-Wish}(\mu_{p \times p}, \nu)$, then, when $\nu > p + 1$, we have $E(X) = \frac{\mu}{\nu - p - 1}$ and $\text{Mode}(X) = \frac{\mu}{\nu + p + 1}$

Therefore, in the previous modelling, we will think of $\theta = F^{-1}(\tilde{\theta})$ where the only parameters which are changed are those who have just been stated above.

4.4 Results

In the first subsection, we will decide on the parameter values which will be used to compare to different models (different priors, covariance function, etc). In all simulations, the predictive posterior distribution was drawn 70,000 times (post burn-in). The x^* state was obtained by using an average of the estimates of A and x_T , from the previous period, and thus $x_* = \frac{1}{N} \sum_{i=1}^N A[i]x_T[i]$. This is a simple and fast way of finding a candidate for x_* . It has some connections to the Kalman prediction formula, since

$$\text{Proj}(x_{T+1} \mid y_{1:T}) = \text{Proj}(A_T x_T \mid y_{1:T-1}) + \text{update term},$$

where $\text{Proj}(x_{T+1} \mid y_{1:T})$ is the best linear predictor of x_{T+1} given the data up to time T , i.e. its projection onto the hyperplane spanned by the data up to time T . When the x_t and $y_{1:T}$ are jointly normally distributed, we have $\text{Proj}(x_{T+1} \mid y_{1:T}) = E(x_{T+1} \mid y_{1:T})$. Therefore, in our Bayesian setting, the formula $x_* = E(A_T x_T) \approx \frac{1}{N} \sum_{i=1}^N A_T[i]x_T[i]$ makes some sense.

4.4.1 A Preliminary Simulation

In this subsection, the data used was comprised of the last 15 observations in Ireland (2004). Starting from 10 observations, we predicted for the 11th period y_t , and then reestimated the model with 11 observations, and so on. We will do 1-lag predictions for the observed values to the right of the vertical dashed red line in Figure 4.2 below.

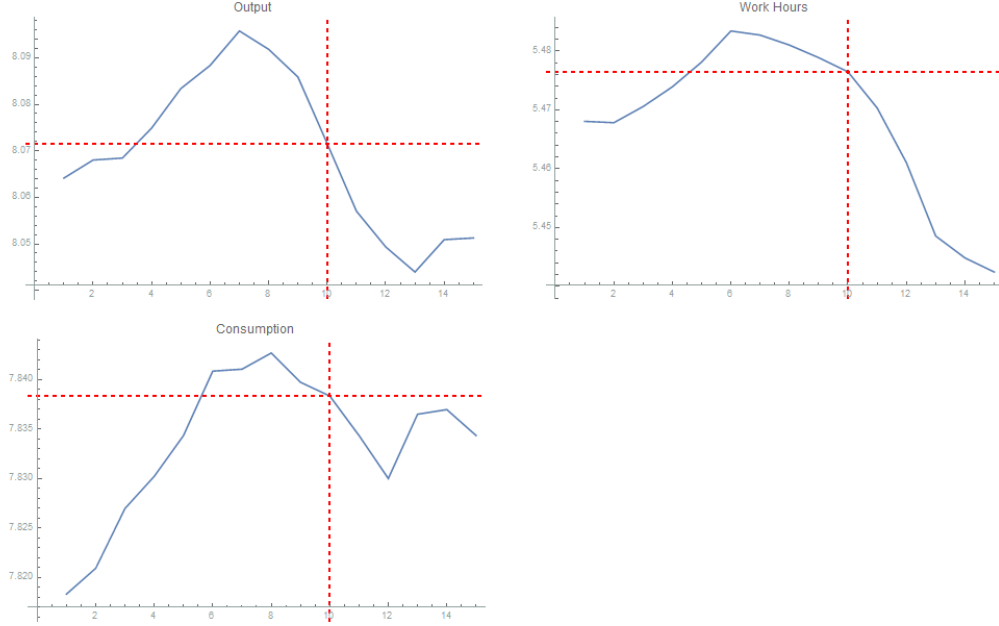


Figure 4.2: Plots for different components of y_t (log-deviations from steady state): 15 last observations

We used 20 particles in the PGAS, a sample size of 55,000, $s_1 = 0.0001$, $s_2 = 0.1$. We used greater, and smaller step sizes, but our simulations returned worse results with respect to the mixing of the chains. We used a burn-in of 5,000. For the predictive a posteriori distribution, we took 100,000. In this simulation, the initial values used were:

$(\beta, \delta, \theta, \rho, A, \gamma) = (0.7, 0.3, 0.6, 0.7, 2, 0.6)$, and $l_{GP} = 20$.

$$M_1 = \begin{bmatrix} 6.528 & 1.065 & 6.565 \\ 1.065 & 1.607 & 0.968 \\ 6.565 & 0.968 & 6.621 \end{bmatrix}, \Sigma = \begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 0.2 & 0 \\ 0 & 0 & 0.3 \end{bmatrix} \text{ and } Q = \begin{bmatrix} 1.5 & 0 \\ 0 & 0.6 \end{bmatrix}$$

It took approximately 62 hours for this simulation to finish, using a server with a CPU of 30 cores. We can observe that extending this simulation to bigger periods will be computationally unbearable specially after noticing that the algorithm has at least $\mathcal{O}(T^3)$ complexity¹³.

Traces

Let us now analyse the traces of our simulation, which may give us clues regarding the mixing of the chains from the algorithm.

¹³See Frigola-Alcalde (2015)

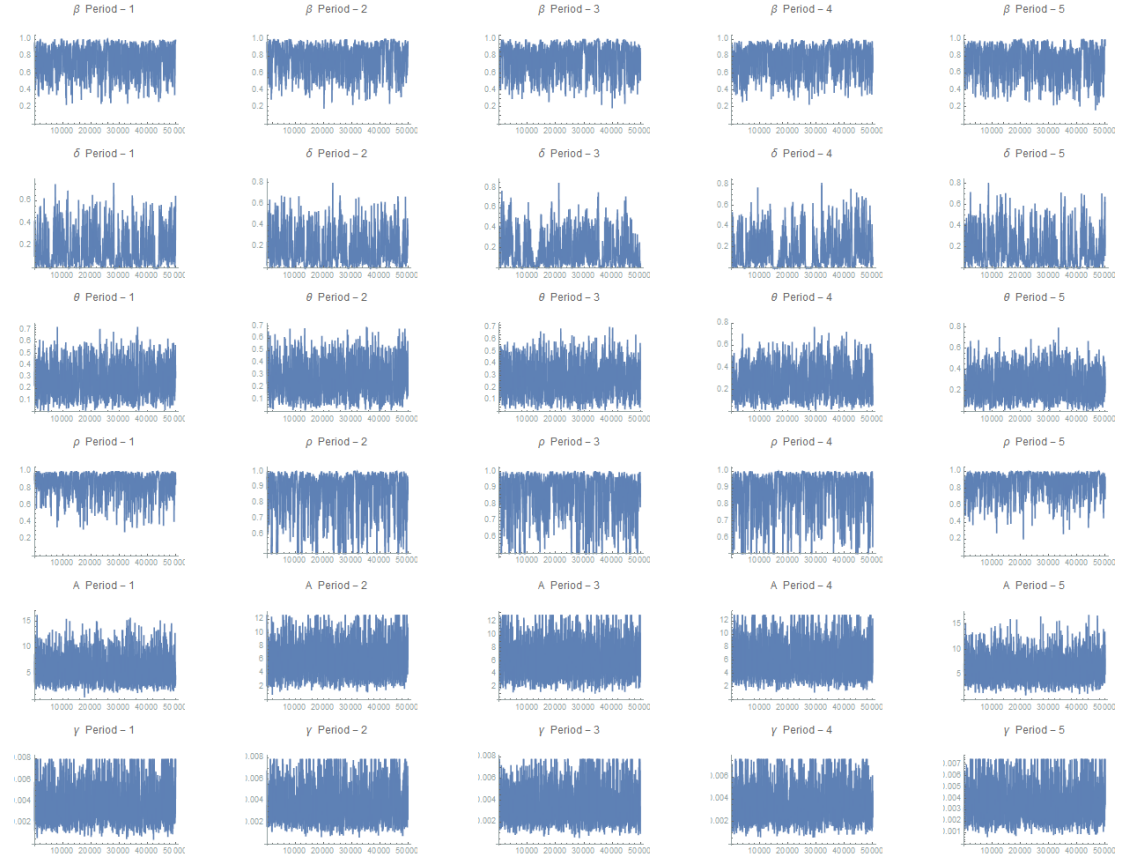


Figure 4.3: Trace of Economic Parameters

There is some room for improvement in the mixing of ρ and δ , but in general the chains mixed reasonably well for all the economic parameters.

Next, we will look at the trace for the M_1 matrix. First, we will use the prior density, evaluated at the chain values, as a univariate indication of how well the chain is mixing, and later we will look at each component of M_1 separately.

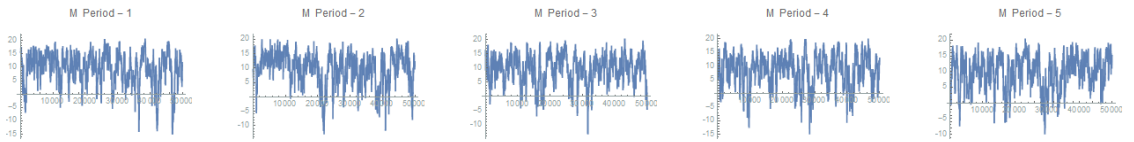


Figure 4.4: Trace of M_1 using its prior density

Looking at each element in particular, the mixing seems to not be too bad, with some clear room for improvement. Also, maybe due to the small sample size, we have an $M1$ indicating possible independent processes.

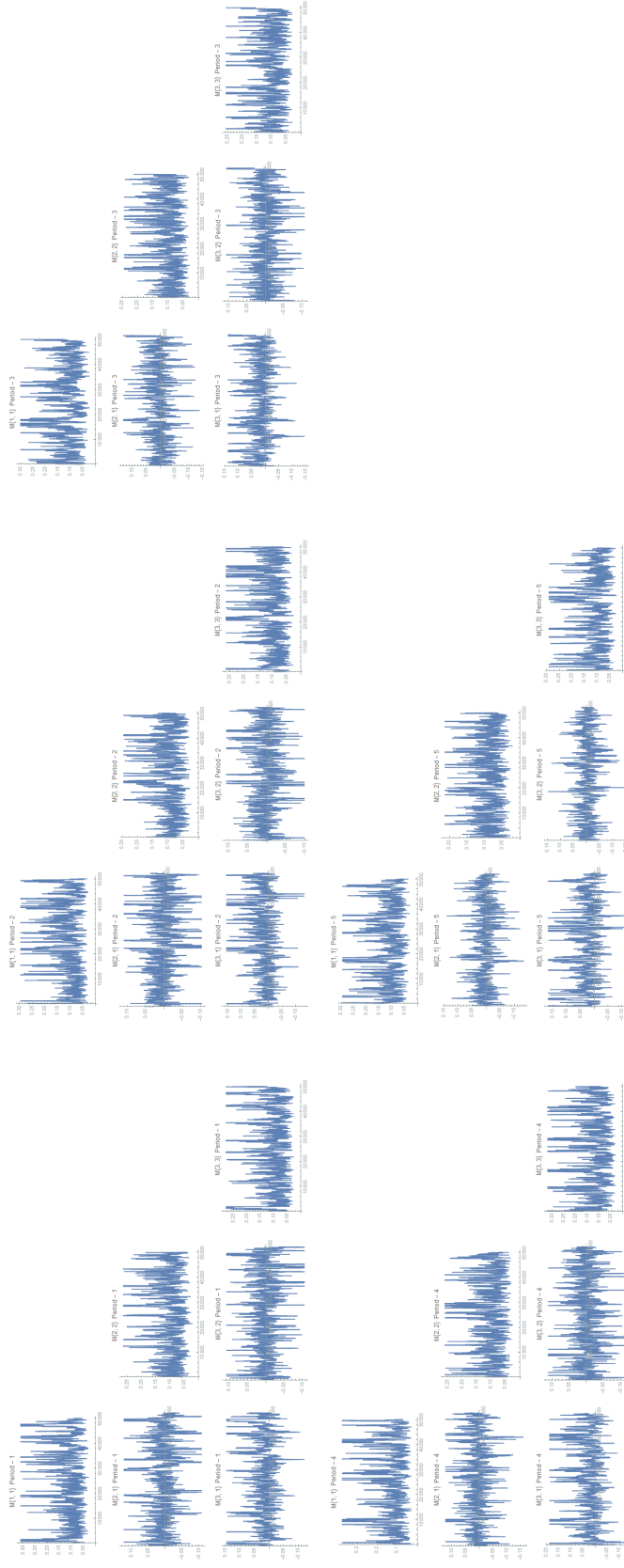


Figure 4.5: Trace of M_1 componentwise



Figure 4.6: Trace of σ_i^2, q_i^2

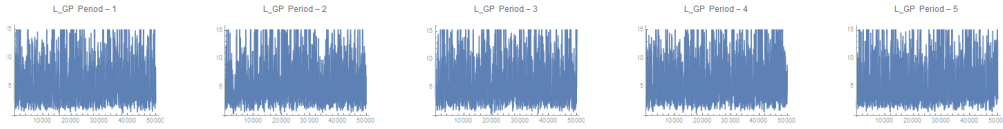


Figure 4.7: Trace of l of the Covariance Function

The traces for the covariance function length parameter, and for the measurement error variances look very good.

Posterior Histograms, and Density Priors

We will now compare the histogram from the estimated posterior distributions with the prior density function. The objective is to observe whether our estimations were somehow influenced by the data, or were they mainly determined by the priors imposed on the parameters. An histogram that differs from the prior density shows that our algorithm learned from the data.

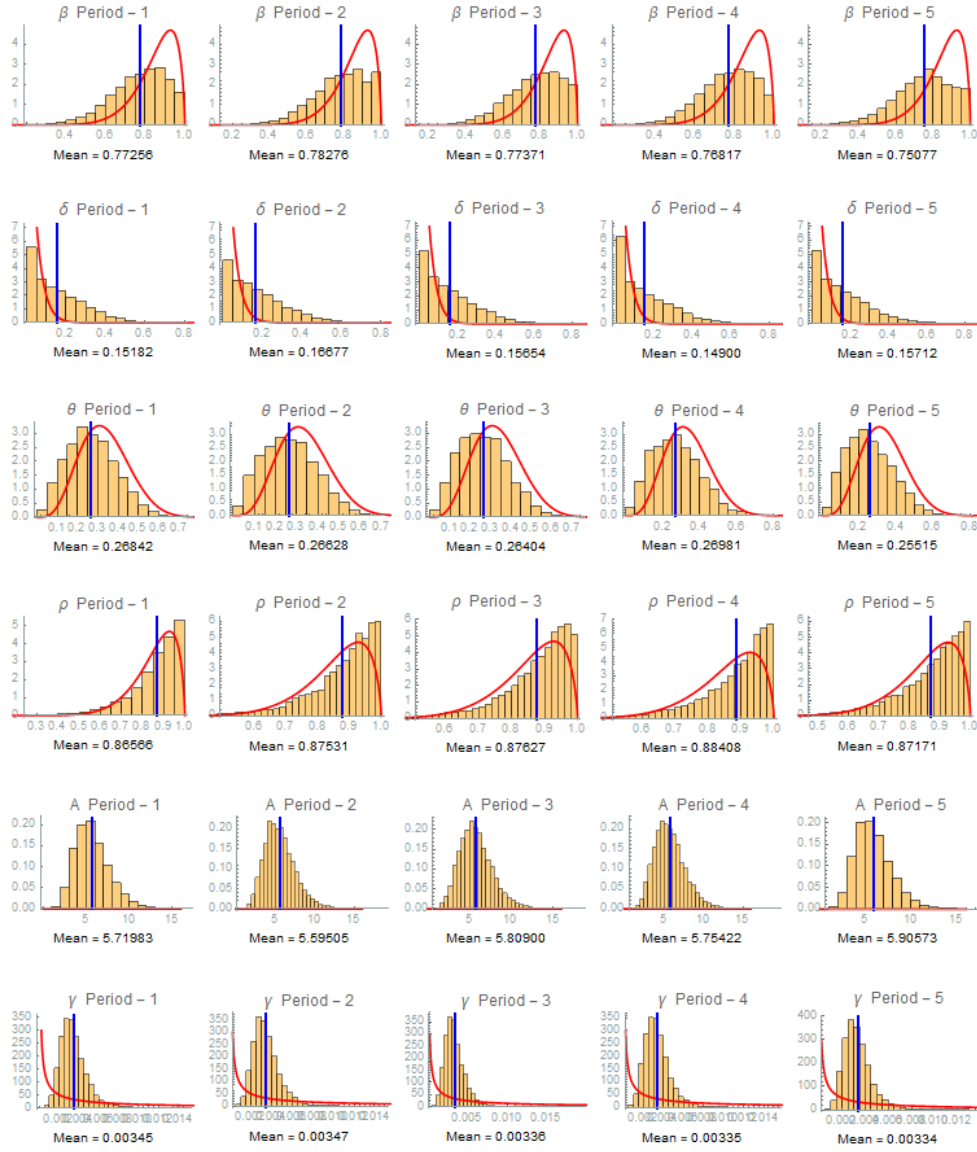


Figure 4.8: Histograms and Density Priors for Economic Parameters

Figure 4.8 reveals that the model learned something across all the economic parameters, even though in some the learning was less pronounced — the exceptions are δ and ρ . The histogram for rho is very close to the prior density, however, we can see a trend of the mode of the posterior distribution getting closer to 1 than what we can observe in the prior. With a greater amount of data, one would expect this behaviour to become more pronounced. The mean estimations are similar to those obtained in Ireland (2004).

The histograms in Figure 4.9 gives us a slightly different information from what we had surmised from the traces for M1 individual components. The prior densities of this plot are not shown, since they live in a space of 7 dimensions, i.e. the density of one element is dependent on the other 6 elements values. Remember that we chose an Inverse Wishart distribution to ensure we had an PD matrix. However, it still does make sense to look at their individual histogram, since there is the possibility the data may have broken that correlation. From the previewed means in Figure 4.9, the magnitude for the M1 component governing the Gaussian Process for Output and Consumption can be $10\times$ higher than Work Hours and any of the remaining components of y_t . We could only find — see I. Alvarez, Niemi, and Simpson (2014) — the marginal distributions for the main diagonal elements of M_1 when the matrix follows an inverse scaled Chi-squared distribution, with the form:

$$\sigma_{ii} \sim \mathcal{I}\chi^2 \left(\nu + d - 1, \frac{\lambda_{ii}}{\nu - d + 1} \right),$$

when $\Sigma \sim \mathcal{IW}(\nu, \Lambda)$. In our case, we have $\nu = 15$, $\Lambda = I_3$ and $d = 3$. To avoid cluttering this section with graphs, and since plotting on Figure 4.9 would require many changes to our code for this Figure, the interested reader may consult the Mathematica files of this work. It suffices to say that the density will look almost like an horizontal red line over the x -axis¹⁴, not so dissimilar to the graphs in 4.9 for parameter A, i.e. there is a very noticeable learning from the data.

¹⁴For the range of x values showed in the Figure 4.9. For greater values, the density will create a somewhat similar wave figure to the χ -squared distribution.

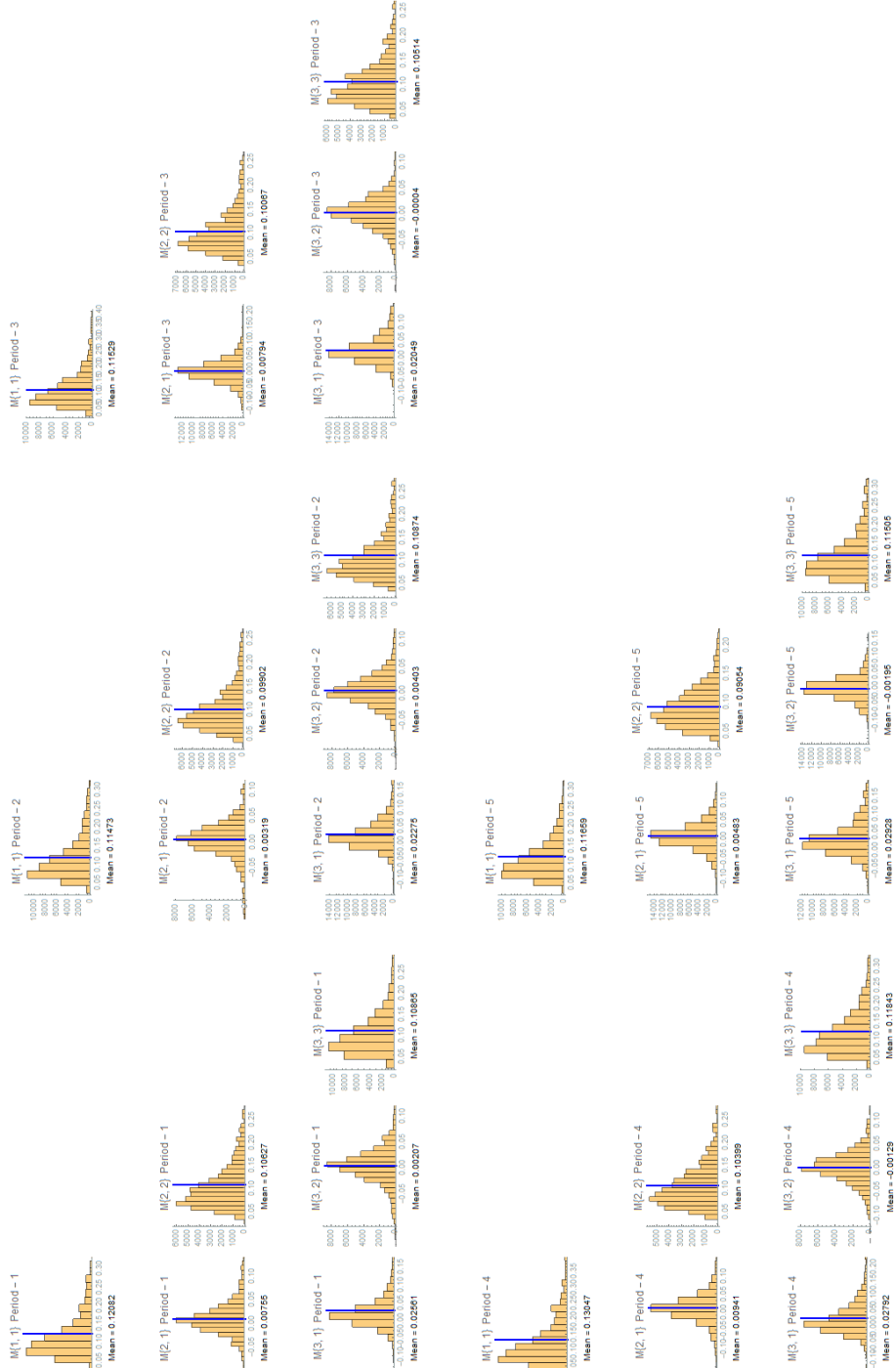


Figure 4.9: Histogram of M_1 componentwise

If we compute the correlation, using the covariances from above, we get:

Period 1	Period 2	Period 3
$\begin{pmatrix} 1 & & \\ 0.0666 & 1 & \\ 0.2235 & 0.0192 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & & \\ 0.0299 & 1 & \\ 0.2037 & 0.0388 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & & \\ 0.0737 & 1 & \\ 0.1861 & -0.0004 & 1 \end{pmatrix}$
Period 4	Period 5	
$\begin{pmatrix} 1 & & \\ 0.0808 & 1 & \\ 0.2246 & -0.0116 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & & \\ 0.0470 & 1 & \\ 0.2528 & -0.0191 & 1 \end{pmatrix}$	

Table 4.3: Correlation matrices for different periods

From Table 4.3, the GP for component 2 (Work Hours) has a very slightly positive correlation with process 1(Output), between 2% and 8%, whereas a near zero correlation with component 3(Consumption). This behaviour for the work hours data may be due to a small sample size, or greater difficulty in the mixing, although this last possibility seems less likely given the traces above.

However, the strongest explanation for this behaviour would be the nature of the data itself. One might be tempted to state that since the bias is unobservable, we cannot substantially conclude something from data. Because in our setting the economic model is not being emulated by a GP, the y^{eco} will act solely as a mean function to the resulting GP of $y^{\text{eco}}(x) + b(x)$, keeping the covariance function of the bias term b responsible for tracking the data correlations.

From economic theory, if output induces changes in (un)employment, usually these are seen at the extensive margin (choosing to work/not work), instead of the intensive margin (increasing/decreasing work hours). Furthermore, Output and Consumption do have a noticeable positive correlation between 18% and 25%, which is something we would again expect from economic theory.

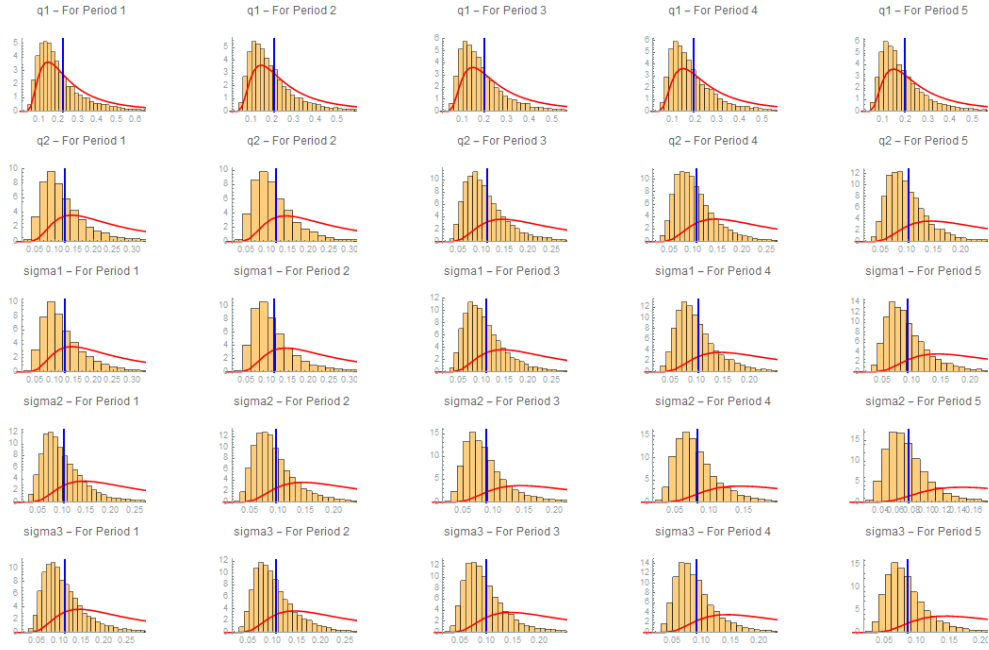


Figure 4.10: Histograms and Density Priors for σ_i^2, q_i^2

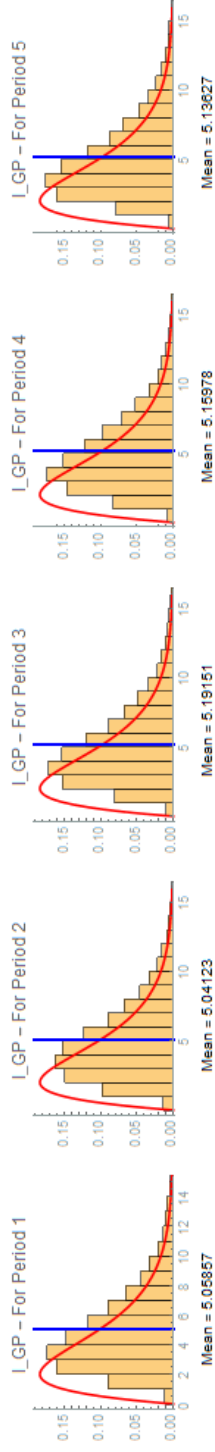


Figure 4.11: Histograms and Density Prior for l of the Covariance Function

In Figure 4.11, one can see an increase in the mean of l , from period 1 to period 2, and to the next period, followed by what seems to be a stabilization afterwards.

82

One-step ahead Predictions

We shall now look at one step ahead (also known as one lag prediction) for our observed variables.

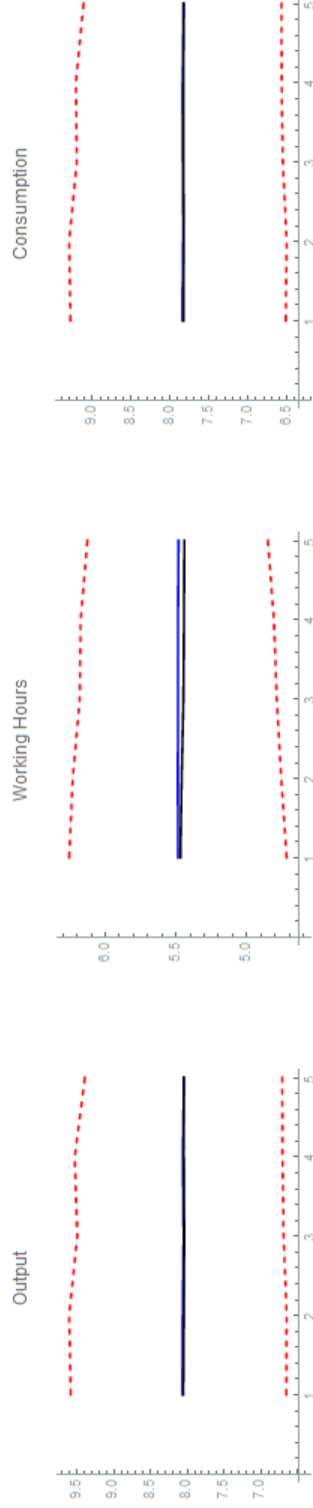


Figure 4.12: Bias-Corrected Predictive plot for 5 periods

Figure 4.12 shows 2 dashed red lines computed from the bias-corrected predictive a posteriori distribution for each component of y_t . They correspond to the 2.5% and the 97.5% percentiles, the black line is the mean prediction, and the blue one is the actual observed value.

The red dashed bars are wide apart, accounting for model uncertainty. Other factors may also contribute to an increase of the uncertainty, when predicting, such as the small sample size. Due to the nature of the macroeconomic science, we have only 1 *noisy* observation for each point in time, contrary to what we would want in other experimental settings. As we increase the sample size, the clear trend is for a tightening of the red lines. However, if one has a great uncertainty associated to their predictions, this could also be an indication of identification issues. We will pursue this topic further in Chapter 6.

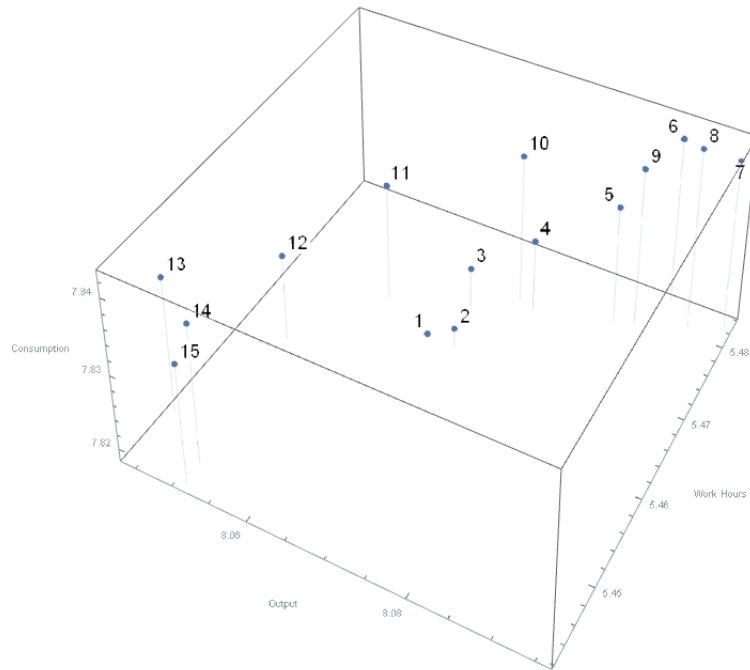


Figure 4.13: 3D representation of the 15 data points

From Figure 4.13, one can see that data point 12 and 13 are farther from previous data, than any other points. So, it may suggest that we could expect a greater uncertainty at either of these points.

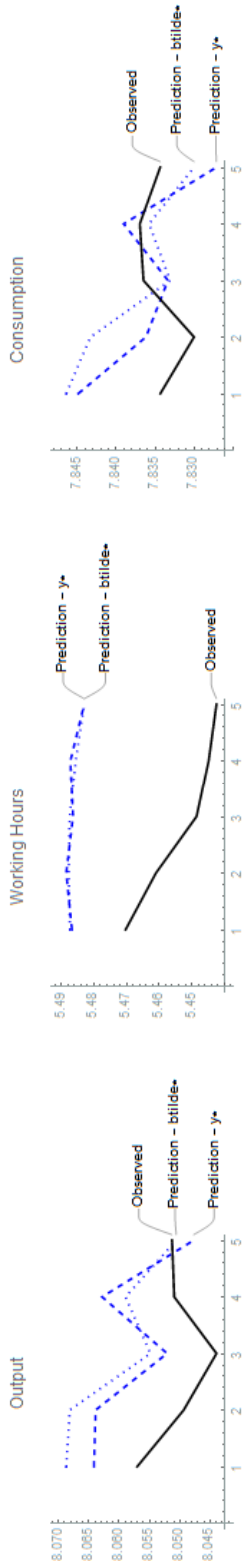


Figure 4.14: Two different ways of prediction

So, far we've used to predict the formula for y_* in section 4.2.2($y^{\text{eco}} + \text{bias} + \text{measurement error}$). However, we could also use \tilde{b}_* ($y^{\text{eco}} + \text{bias}$) in the same section.

In Figure 4.14 we show both. We could also expect the uncertainty red bars for \tilde{b}_* to be tighter, and in fact it is what happens in Figure 4.15

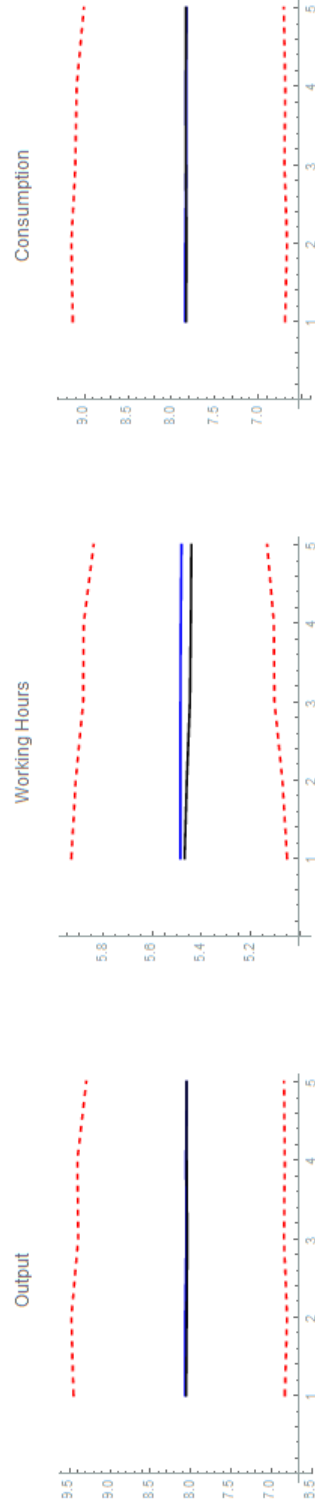


Figure 4.15: \tilde{b}_* Prediction uncertainty bars

The measurement error, if there is no identification issue, seems to have a negligible effect. Looking at the 90% prediction bands of Figure 4.15, we still see what seems to be considerably large interval with respect to the movement of the series¹⁵. One possible explanation for this could also be the very diffuse prior which were used for this exercise. One could also think that it could be the case of being in the presence of an identification issue, such as those referred previously, at the beginning of this work. In Chapter 6 of this work, we will dwell more on this possibility, and simulate a forecasting exercise with several different priors.

¹⁵We could not find a suitable criteria for the justification of this statement other than observing the straightness of the observed and predicted 'irregular' lines. However, given the time period being analysed — the dot com bubble — we expect the movements around the steady-state to be greater than those in 'normal' times, possibly entailing an increased uncertainty when predicting.

To analyse just how much the inclusion of the bias term has improved the forecasting of the model, we must compare the bias-corrected model to the original model estimated with no bias term, henceforth denoted by y_M . For details on how the computation was done, we refer the reader to Appendix B.3.

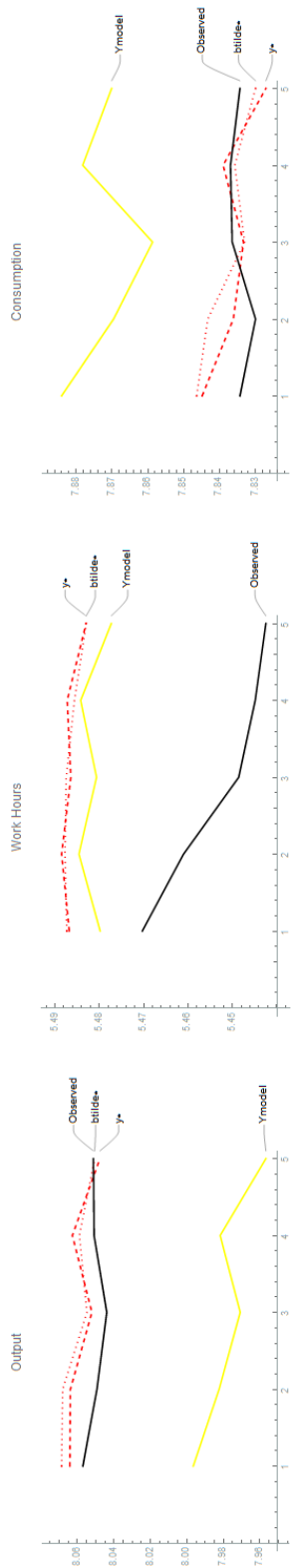


Figure 4.16: Predictions from y_M , \tilde{b}_* and y_*

From 4.16 the improvement is great for Output and Consumption, namely the processes which were positively correlated. However, for Work Hours the performance was very similar with respect to the pure model. A possible explanation can be brought to light, if we look at Figure 4.2. We can see that the periods to be predicted are to the right of the vertical dashed red line. Only for Consumption are those periods inside the range of values which were used to estimate the model. So, for this component we may expect our predictions to act as interpolation. Although for Output, the period values are outside the range, Output is very correlated with Consumption, which may help with the predictions for this component. However, Work Hours is near independent from the remaining components, and also the observed values for these prediction periods are farther away from those used for estimation than what we observe for the other components.

Hence, for Work Hours, the prediction exercise is more similar to an extrapolation exercise, and in these the Gaussian Processes have severe known limitations. A possible solution may be to increase the sample size, and make predicting more similar to interpolation.

For a quantitative analysis, in our 1-lag prediction for the 5 periods, using the y_* formula, we get a RMSE for each y component of

$$(0.00981434, 0.0343616, 0.0065688)$$

If we use the \tilde{b}_* formula we have

$$(0.0115895, 0.0341235, 0.00842655)$$

For the y_M predictions, we got $(0.0737338, 0.0298013, 0.0387475)$. Comparing to y_* we get a factor of $(7.51286, 0.867285, 5.89871)$, and to \tilde{b}_* we get a factor of $(6.3621, 0.873339, 4.59826)$.

4.4.2 Comparing Performances

A simulation exercise which is done in Ireland (2004) is to predict the last 69 observations for $y = (\text{Output}, \text{Work Hours}, \text{Consumption})$. In this section, we shall use our method to forecast for the same periods. A few changes are in order for to make the MCMC tries in this paper to be more amenable, since the estimation and prediction using our model are computationally quite intensive. Instead of learning from an increasing set of observations, our model will learn only from the last 10 previous observations, and the initial values for the parameters will be the average of the previous period estimation. To make this exercise faster, we only used an MCMC chain of size 15000, with a size burnin of 5000. This exercise lasted approximately 8 days and 18h.

Proceeding in this way, we get a RMSE for y of

$$(0.0148201, 0.0182601, 0.0157189)$$

This is an improvement, when comparing with Ireland (2004) RMSE results for 1 period ahead forecast of $(0.8319, 0.5371, 0.5345)$. However, the forecasting exercise in Ireland (2004) was done by detrending the data in each of the periods, i.e. re-estimating η in each period, and keeping β and δ fixed. To preclude our improvement of performance to be solely due to these different aspects, in the next section we will redo this exercise by detrending data and with those two parameters fixed.

Detrending Data, Fixing β and δ

To make our comparison to Ireland (2004) results fairer, we now proceed by keeping β and δ fixed at 0.99 and .025 respectively.

For the detrending of data, since our UQ methodology only deals with time-invariant systems directly, we must detrend it before learning the UQ model. We will estimate η up to period T , then in the next estimation period, we use the data up to period $T + 1$ data.

To estimate this parameter we use the definition of $y_t = \frac{Y_t}{\eta^t}$ and so, we are expecting the following relationship:

$$\log(Y_t) = t \log(\eta) + \log(y_t)$$

So, we regress all previous log output data on t in a linear model of the form

$$\log(Y_t) = \beta_0 + t\beta_1 + \epsilon_t$$

Proceeding in this way, we get a RMSE for y of

$$(0.0918451, 0.0194998, 0.040572)$$

If we analyse these results, although in line with Ireland (2004), the errors are much smaller, at least one order of magnitude smaller. We still keep the same relationship between them, with the Output component being the hardest to forecast, followed by Consumption, and then Work Hours. These last modifications to our forecasting exercise did not significantly change the RMSE for Work Hours, which was to be expected, because that component is not influenced by the time trend, i.e. η re-estimation does not affect it.

4.5 Conclusions

In this chapter we have used a full gaussian process model discrepancy term with a dual objective. First and foremost, we tried to account for model misspecification in a DSGE model, in its state-space representation, and secondly to improve on its parameters estimation and its forecast performance. On our second objective, we were only partially successful, since the learning of our model was very similar to that which we would obtain while using the Ireland approach. The posterior distributions obtained for the parameters could indicate a possible identification issue, although it is not clear, since we are using very diffuse priors, and a small number observations. Further simulations are warranted, namely with tighter priors, and a deeper understanding on the resulting posteriors. An optimal outcome would be a

considerable tightening of the posteriors. Despite this possible issue, our predictive performance does not appear to suffer from it, as we can observe from the RMSE of the above simulations, and in other situations such as Arendt, Apley, and Chen (2012). Until now, and to our knowledge, techniques for correcting our identification issues in GP regression, such as in Plumlee and V. Joseph (2018), are too limitative and impose further computational costs to an already heavily burdened method, which is the greatest shortcoming of our UQ method.

Venues for future research would be improving the computational efficiency, either by using approximations to our GP, or using online algorithms such as in M. Cai et al. (2019), to use a lower level programming language, or to improve the mixing of the Markov Chains for the parameters draws. Another venue would be to further methods to correct identification issues, while keeping computational tractability.

Chapter 5

UQ using a Hilbert Reduced Rank Approximation to a GP

5.1 Introduction

In the previous chapter, we used a model bias term with a full Gaussian Process prior as an attempt to correct for misspecification while preserving the scientific interpretability of the estimated parameters. However, the resulting methodology was computationally very demanding. In this work, we shall use instead the functional approximation to the Gaussian Process from Solin and Särkkä (2014) and Svensson et al. (2016), with the purpose of studying model discrepancy without those shackles. From what the literature has produced, we can expect a much faster computation time, but with the downside of an increase, hopefully small, in the predictions RMSE.

As shown in Solin and Särkkä (2014), associated to each covariance function $k(x, x')$, we can also define a covariance operator \mathcal{K} as $\mathcal{K}\phi = \int k(\cdot, x')\phi(x') dx'$. With this interpretation, we could apply an approximation to the covariance operator using what is known as Hilbert Space methods, usually used for approximating differential and pseudo-differential operators when we are dealing with partial differential equations. This method also has connections with Sturm-Liouville theory.

Perhaps the easiest method to derive the approximation is the *inner product* perspective developed in Solin and Särkkä (2014), which we will follow somewhat closely. For more information on the following deductions, the interested reader may consult this same article, and the references therein.

Let us define the inner product by

$$\langle f, g \rangle = \int f(x)g(x)w(x) dx$$

with $w(x)$ being a positive integrable weight function. Defining the operator

$$\mathcal{K}f = \int k(\cdot, x)f(x)w(x) dx,$$

this operator is self-adjoint, and hence normal. Therefore, by the Spectral Theorem for compact normal operators¹, there exists an orthonormal set of basis functions $\phi_j(x)$ and positive constants γ_j , such that

$$k(x, x') = \sum_j \gamma_j \phi_j(x) \phi_j(x')$$

and it can be shown that $\gamma_j = S(\sqrt{\lambda_j})$, where $S(\cdot)$ is the spectral density of the covariance function.

Using the Karhunen-Loève (see section C.1.2 in Appendix C of this text for further information and references) expansion for a stochastic function $f(x)$ with zero mean and k as covariance functions, we are able to write

$$f(x) = \sum_j f_j \phi_j(x)$$

where f_j are independently distributed $N(0, S(\sqrt{\lambda_j}))$.

5.1.1 Hilbert Space Methods for a Reduced-Rank Multidimensional GP

Part of the computational limitations we encountered on Chapter 4, are due to the use of a full GP as the prior for the bias term. This computational burden, when using a full GP, has already been seen in different settings. The interested reader may find some examples in Solin and Särkkä (2014). Several methods to improve the efficiency of its use have been proposed. Those present in this section can be considered to pertain to a class of methods called *reduced-rank* approximations. These methods are characterised by approximating the covariance matrix $K(X, X)$ by another one with a smaller rank, allowing for the use of the Woodbury formula, giving a fast approximation to the inverse of the matrix. The method proposed in Solin and Särkkä (2014) takes advantage of an approximate eigen-decomposition of covariance functions with respect to an eigenfunction expansion of the Laplace operator in a compact subset of \mathbb{R}^{D_x} ². In the context under consideration, we will

¹See Appendix C

²The exact formalization of the eigenvalue problem for the Laplace operators with Dirichlet boundary conditions is the following: Let $\Omega \in \mathbb{R}^{D_x}$ and consider $-\nabla^2 \phi_i(x) = \lambda_i \phi_i(x)$ when $x \in \Omega$, and $\phi_i(x) = 0$ when $x \in \partial\Omega$. One should also keep in mind that $-\nabla^2$ is a positive definite Hermitian operator.

have multidimensional output and input. Assuming that the D_x dimensional inputs live in a rectangular domain $\Omega = [-L_1, L_1] \times \cdots \times [-L_{D_x}, L_{D_x}]$, with Dirichlet boundary conditions, i.e. with conditions on the values the solution to a partial differential equation must take on the boundary, and considering an approximation truncated at degree $m = (m_1, \dots, m_{D_x})$, we have:

$$\tilde{k}_m(x, x') \approx \sum_{(j_1, \dots, j_{D_x})=\mathbf{1}}^m S(\sqrt{\lambda_{j_1, \dots, j_{D_x}}}) \phi_{j_1, \dots, j_{D_x}}(x) \phi_{j_1, \dots, j_{D_x}}(x'),$$

where $S(\cdot)$ is the spectral density of the covariance function, with eigenfunctions of the Laplace operator, in a fixed domain, given by

$$\phi_{j_1, \dots, j_{D_x}}(x) = \prod_{k=1}^{D_x} \frac{1}{\sqrt{L_k}} \sin\left(\frac{\pi j_k (x_k + L_k)}{2L_k}\right),$$

and with eigenvalues

$$\lambda_{j_1, \dots, j_{D_x}} = \sum_{k=1}^{D_x} \left(\frac{\pi j_k}{2L_k}\right)^2.$$

To use the above notation in a multidimensional context, and still use matrices, we need an injective function, mapping (j_1, \dots, j_{D_x}) into \mathbb{N} , and one such intuitive map is the base- m expansion of integers, which is given as follows:

$$\mathcal{E}(j_1, j_2, \dots, j_{D_x}) = 1 + \sum_{k=1}^{D_x-1} (j_k - 1) m_{k+1} m_{k+2} \cdots m_{D_x} + (j_{D_x} - 1),$$

and therefore, we can now identify $j = (j_1, j_2, \dots, j_{D_x})$ with $\mathcal{E}(j_1, j_2, \dots, j_{D_x})$.

Assuming for the moment that we are dealing with the unidimensional output case, the linear regression model is

$$\begin{aligned} f &\sim \mathcal{GP}(0, k(.,.)) \\ y_t &= f(x_t) + \epsilon_t, \end{aligned}$$

with $\epsilon_t \sim N(0, \Sigma)$ and from using the Karhunen-Loève expansion, it will result that³

$$f(x_t) \approx \sum_{j=1}^m w_j \phi^{(j)}(x_t),$$

³The notation is $j \cong \mathcal{E}(j)$

and for the weights above, we would define $w_j \sim \mathcal{N}(0, S(\sqrt{\lambda_j}))$, where $S(\cdot)$ is the spectral density of the covariance function of the Gaussian Process.

A way to generalize to the multidimensional output case is to consider

$$f(x_t) \approx W\phi(x_t),$$

where $\phi(x_t) = \begin{bmatrix} \phi_{(1,\dots,1)}(x_t) \\ \vdots \\ \phi_{(m_1, m_2, \dots, m_{D_x})}(x_t) \end{bmatrix}$ ⁴ and $W = \begin{bmatrix} w_1^{(1)} & \dots & w_{\mathcal{E}(m)}^{(1)} \\ \vdots & & \vdots \\ w_1^{(D_y)} & \dots & w_{\mathcal{E}(m)}^{(D_y)} \end{bmatrix}$ with dimensions $D_y \times \mathcal{E}(m)$, and impose on W the following prior conditioned on Σ :⁵

$$W \mid \Sigma \sim \text{Matrix-Normal}(0, \Sigma, \text{Diag}(S(\sqrt{\lambda_j}))),$$

with mean zero, row covariance matrix Σ , and column covariance matrix $\text{Diag}(S(\lambda_j))$. It is important to notice that the vector $\phi(\cdot)_{\mathcal{E}(m) \times 1}$ is deterministic, and hence independent from the model parameters.

5.1.2 Hilbert Reduced-Rank GP in a State-Space Model with a Discrepancy Term

In our SSM, and following the model discrepancy framework presented previously, in Chapter 4, together with the notation used in the subsection 5.1.1, we have

$$\begin{aligned} x_{t+1} &= Ax_t + \eta_t \\ y_t &= C_* + Cx_t + b(x_t) + \epsilon_t, \end{aligned}$$

where $\eta_t \sim N(0, Q)$, and $\epsilon_t \sim N(0, \Sigma)$ and $b(x_t) \sim \mathcal{GP}(0, k_M(\cdot, \cdot))$.

Doing a m -degree approximation, with the priors in Svensson et al. (2016), to the mean zero GP prior, we obtain

$$b(x_t) \approx W\phi(x_t)$$

$$W \mid \Sigma \sim \text{Matrix-Normal}(0, \Sigma, V)$$

with $V = \text{Diag}(S_M(\sqrt{\lambda_j}))$ and

$$\Sigma \sim \text{Inv-Wishart}(l_\Sigma, \Lambda_\Sigma).$$

⁴Even though the order in which each index changes may be subjective, we must be careful to be consistent through out the whole algorithm

⁵See Appendix B

To use the expansion, we must reformulate the observation equation such that

$$\tilde{y}_t := y_t - (C_* + Cx_t) = b(x_t) + \epsilon_t$$

and adapting the formulas from Svensson et al. (2016) to our context, we get the following conditional distributions⁶

$$W|x_{1:T}, y_{1:T}, \Sigma \sim \mathcal{MN}\left(\tilde{\Phi}(\tilde{\Psi} + V^{-1})^{-1}, \Sigma, (\tilde{\Psi} + V^{-1})^{-1}\right),$$

where $\tilde{\Phi} = \sum_{t=1}^T \tilde{y}_t \phi(x_t)^\top$, $\tilde{\Psi} = \sum_{t=1}^T \phi(x_t) \phi(x_t)^\top$, and

$$\Sigma|x_{1:T}, y_{1:T} \sim \text{Inv-Wishart}\left(T + l_\Sigma, \Lambda_\Sigma + \tilde{\Upsilon} - \tilde{\Phi}(\tilde{\Psi} + V^{-1})^{-1}\tilde{\Phi}^\top\right),$$

with $\tilde{\Upsilon} = \sum_{t=1}^T \tilde{y}_t \tilde{y}_t^\top$.

To keep the notation simple, we have omitted the fact that the conditional distribution for W and Σ depend on more parameters, not explicit on the notation.

One should notice, that although in the deduction above $\tilde{y}_t = y_t - g(x_t)$ where $g(x_t)$ is linear, we could apply our method to any type of non-autonomous function of x_t , not necessarily linear. We use a linear function, since the model in Ireland (2004) is a linear model. Nothing prevents the application of this method to more complex, non-linear SSMs.

5.1.3 Prediction in a SSM with a HRRGP Prior on a Discrepancy Term

Similarly to the previous chapter, we will use the model from Ireland (2004), before the addition of the VAR term, with the same modifications as before, i.e.

$$x_t = Ax_{t-1} + \eta_t,$$

where $\eta_t \stackrel{iid}{\sim} N(0, Q)$, with Q assumed diagonal. Furthermore, instead of a bias term with a full GP prior in the observation equation, we add an m -degree approximation of the bias term, which has a GP prior, to the observation equation to measure model bias/misspecification

$$y_t = C_* + Cx_t + W\phi(x_t) + e_t,$$

where W has a Matrix-Normal distribution, and $e_t \sim N(0, \Sigma)$, where Σ is allowed to be a non-diagonal symmetric and positive definite matrix.

⁶There are some minor typos in the paper. See Appendix B for a deduction of the formulas for the Conjugate Prior

To make predictions for an arbitrary x_* , and $\theta = (W, \Sigma, \theta_{-W\Sigma})$ being the parameters of the SSM, we have the following Monte Carlo approximation:

$$\begin{aligned} p(y_* \mid x_*, y_{1:T}) &= \int p(y_* \mid x_*, y_{1:T}, \theta) p(\theta, x_{1:T} \mid x_*, y_{1:T}) d\theta dx_{1:T} \\ &\approx \frac{1}{N} \sum_{i=1}^N p(y_* \mid C_*[i], C[i], W_*[i], x_*, \Sigma[i]) \\ &= \frac{1}{N} \sum_{i=1}^N N(y_* \mid C_*[i] + C[i]x_* + W_*[i]\phi(x_*), \Sigma[i]), \end{aligned}$$

where to sample from $p(\theta, x_{1:T} \mid x_*, y_{1:T})$, we shall use the PGAS algorithm from Lindsten, Jordan, and Schön (2014).

From the exposition in Lindsten, Jordan, and Schön (2014), we have **Algorithm 1** below.

Algorithm 4 PGAS for Bayesian Learning of SSMs

- 1: Set $\theta[0]$ and $x_{1:T}[0]$ from some distribution
 - 2: **for** $n \geq 1$ **do**
 - 3: Draw $\theta[n] \sim p(\theta \mid x_{1:T}[n-1], y_{1:T}, \theta[n-1])$ /* Using Algorithm 5
 - 4: Draw $x_{1:T}[n] \sim p_{\theta[n]}^N(x_{1:T}[n-1], \cdot)$ /* Using Algorithm 7 */
 - 5: **end for**
-

5.1.4 Sampling from the Parameters

We can sample from the parameters distribution $p(\theta \mid x_{1:T}[n], y_{1:T})$, using a Metropolis-Hastings (M-H) step, together with posterior conditional distributions derived above.

Algorithm 5 Learning Step

Input: $x_{1:T}[n], y_{1:T}, \theta[n-1]$

Output: $\theta_{-W\Sigma}[n], \Sigma[n], W[n]$

- 1: Draw $\Sigma[n] \mid x_{1:T}[n], \theta_{-W\Sigma}[n-1], y_{1:T}$ from

$$\text{Inv-Wishart}(\Sigma \mid T + l_\Sigma, \Lambda_\Sigma + \tilde{\Upsilon} - \tilde{\Phi}(\tilde{\Psi} + V^{-1})^{-1}\tilde{\Phi}^\top)$$

- 2: Draw $W[n] \mid x_{1:T}[n], \Sigma[n], \theta_{-W\Sigma}[n-1], y_{1:T}$

$$\text{Matrix-Normal}(\tilde{\Phi}(\tilde{\Psi} + V^{-1})^{-1}, \Sigma, (\tilde{\Psi} + V^{-1})^{-1})$$

- 3: Draw $\theta_{-W\Sigma}[n] \mid x_{1:T}[n], W[n], \Sigma[n], y_{1:T}, \theta_{-W\Sigma}[n-1]$ /* Use Algorithm 6
-

To use the Metropolis-Hastings step, we will need to deduce the likelihood function for our model. Let us define $\theta = (\theta_x, \theta_y)$, in such a way that makes the next computation possible:

$$\begin{aligned}
p(x_{1:T}, y_{1:T} \mid \theta) &= p(x_1) \prod_{t=1}^{T-1} p(x_{t+1} \mid x_t, \theta_x) \cdot \prod_{t=1}^T p(y_t \mid x_t, \theta_y) \\
&= p(x_1) \prod_{t=1}^{T-1} N(x_{t+1} \mid Ax_t, Q) \cdot \prod_{t=1}^T N(y_t \mid W\phi(x_t), \Sigma) \\
&= p(x_1) N(x_{2:T} \mid (I_{T-1} \otimes A)x_{1:T-1}, I_{T-1} \otimes Q) \\
&\quad \cdot N(y_{1:T} \mid (I_T \otimes C)x_{1:T} + C_{*1:T} + (I_T \otimes W)\phi_{1:T}, I_T \otimes \Sigma)
\end{aligned}$$

$$\text{where } \phi_{1:T} := \begin{bmatrix} \phi(x_1) \\ \vdots \\ \phi(x_T) \end{bmatrix}$$

We will use a MH step for each block $i \in \{1, \dots, B\}$, but building on the previous iteration, i.e. we will consider each time the following vector parameters estimation: $(\theta_{-W\Sigma, < i}[n], \theta_{-W\Sigma, i}^*, \theta_{-W\Sigma, > i}[n-1])$, with $\theta_{-W\Sigma, i}^*$ being the new draw for block i .

Algorithm 6 Metropolis-Hastings Step

Input: $x_{1:T}[n], W[n], \Sigma[n], y_{1:T}, \theta_{-W\Sigma}[n-1]$ **Output:** $\theta_{-W\Sigma}[n]$

- 1: **for** block $i \in \{1, \dots, B\}$ **do**
- 2: Draw $\theta_{-W\Sigma,i}^*$ from their respective proposal density
 $q_i(\theta_i^* | \theta_{<i}[n], \theta_i^*, \theta_{>i}[n-1], x, y)$
- 3: Define $A = \min(1, f)$ with

$$f = \frac{p(x, y | \theta_{-W\Sigma, <i}[n], \theta_{-W\Sigma,i}^*, \theta_{>i}[n-1]) \cdot p(\theta_{-W\Sigma, <i}[n], \theta_i^*, \theta_{-W\Sigma, >i}[n-1])}{p(x, y | \theta_{-W\Sigma, <i}[n], \theta_{-W\Sigma, \geq i}[n-1]) \cdot p(\theta_{-W\Sigma, <i}[n], \theta_{-W\Sigma, \geq i}[n-1])} \cdot \frac{q(\theta_{-W\Sigma,i}[n-1] | \theta_{-W\Sigma, <i}[n], \theta_{-W\Sigma,i}^*, \theta_{-W\Sigma, >i}[n-1], x, y)}{q(\theta_{-W\Sigma,i}^* | \theta_{-W\Sigma, <i}[n], \theta_{-W\Sigma, \geq i}[n-1], x, y)}$$

where, we used the fact

$$p(\theta_{-W\Sigma} | x_{1:T}, W[n], \Sigma[n], y_{1:T}) \propto p(x_{1:T}, y_{1:T} | \theta_{-W\Sigma}, W[n], \Sigma[n]) \cdot p(\theta_{-W\Sigma})$$

- 4: Draw $u_i \sim \text{Unif}[0, 1]$
 - 5: **if** $u_i \leq A$ **then**
 - 6: $\theta_i[n] = \theta_i^*$
 - 7: **else**
 - 8: $\theta_i[n] = \theta_i[n-1]$
 - 9: **end if**
 - 10: **end for**
-

In our implementation of this algorithm we found that some computed quantities were beyond machine-precision. Hence, we applied $\log(\cdot)$ function when computing the acceptance probability of the MH-Step. There is some precedence in its usage, as is evidenced by Cappé, Godsill, and Moulines (2007).

5.1.5 Sampling from the State-Space

If we apply the PGAS Markov Kernel of Lindsten, Jordan, and Schön (2014) to our context, we will obtain the following **Algorithm 7** below. ⁷

⁷For a deduction of the formulas, see Appendix B.

Algorithm 7 PGAS Markov Kernel

Input: $x_{1:T}[n-1]$, $\theta[n-1]$, and N_p which is the number of particles

Output: $x_{1:T}[n]$

- 1: Set $\tilde{x}_{1:T} = x_{1:T}[n-1]$ as the reference trajectory
 - 2: Draw $x_1^i \sim p(x_1|\theta[n-1])$ for $i = 1, \dots, N_p - 1$.
 - 3: Set $x_1^{N_p} = \tilde{x}_1$
 - 4: Set $w_1^i = \frac{1}{N_p}$ for $i = 1, \dots, N_p - 1$.
 - 5: **for** $t = 2, \dots, T$ **do**
 - 6: Draw a_t^i with $P(a_t^i = j) \propto w_{t-1}^j$ for $i = 1, \dots, N_p - 1$.
 - 7: Draw $x_t^i \sim N(A[n-1]x_{t-1}^{a_t^i}, Q[n-1])$ for $i = 1, \dots, N_p - 1$.
 - 8: Compute $\{\tilde{w}_{t-1|T}^i\}_{i=1}^{N_p}$ using
$$\tilde{w}_{t-1|T}^i \propto w_{t-1}^i N(\tilde{x}_t \mid A[n-1]x_{t-1}^i, Q[n-1])$$
 - 9: Draw $a_t^{N_p}$ with $P(a_t^{N_p} = j) \propto \tilde{w}_{t-1|T}^j$
 - 10: Set $x_t^{N_p} = \tilde{x}_t$
 - 11: Set $x_{1:t}^i = (x_{1:t-1}^{a_t^i}, x_t^i)$ for $i = 1, \dots, N_p$.
 - 12: Compute $w_t^i = N(y_t \mid C_*[n-1] + C[n-1]x_t^i + W[n-1]\phi(x_t^i), \Sigma[n-1])$
for $i = 1, \dots, N_p$.
 - 13: **end for**
 - 14: Sample k with $P(k = i) \propto w_T^i$ and set $x_{1:T}[n] = x_{1:T}^k$
-

Similarly to the MH-step, the computed weights took smaller values than machine precision, and so we had to first transform the values using log, then divide by a suitable constant through all weights, so that the normalized values would respect machine precision. This is also another recommendation from Cappé, Godsill, and Moulines (2007).

5.2 Covariance Function, Priors and Proposal

5.2.1 Covariance Function

The Covariance Function to be chosen shall be of the Matérn Class. Besides the reasons alluded in the previous chapter, the objective of choosing the same class is to also allow for a comparison with the results previously presented. For the reader's convenience, we will repeat the characterisation of that class in this chapter. Matérn Class of functions is given by

$$k_M(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{r\sqrt{2\nu}}{l} \right)^\nu K_\nu \left(\frac{r\sqrt{2\nu}}{l} \right)$$

where $r = \|x - x'\|$. The function K_ν is the Modified Bessel of the 2nd kind. When $r = 0$, the right limit of $k_M(r)$ is 1. For fixed r and ν , k_M is increasing in l , and so l is known as the range or length-scale parameter. If we want more spread observations to be more correlated, we use a greater value of l . Both ν and l are positive reals. For this class of functions the Gaussian Process is k -times differentiable if and only if $\nu > k$. For $\nu \rightarrow \infty$ we obtain the squared exponential function. For this reason ν is known as the smoothness parameter. As before, we will decide for rougher processes, so as to allow for an increased richness in the function space. Therefore, similarly we will fix $\nu = 0.5$, as is usual in the literature for very rough processes, for which the Matérn function form simplifies to

$$k_M(r) = \exp^{-r/l}$$

We will estimate from the data the value for the length-scale parameter.

Fortunately for our work, since we chose the Matérn kernel, we may easily consult the respective spectral density:

$$S_M(s) := \frac{2^{D_x} \pi^{\frac{D_x}{2}} \Gamma(\nu + D_x/2) (2\nu)^\nu}{\Gamma(\nu) l^{2\nu}} \left(\frac{2\nu}{l^2} + 4\pi^2 s^2 \right)^{-(\nu + D_x/2)}$$

5.2.2 Priors

The macroeconomic common practice is to put priors directly on the economic parameters, we will consider jointly the θ_x and θ_y , with

$$p(\theta | x_{0:T}[n], y_{1:T}) \propto p(x_{0:T}[n], y_{1:T} | \theta) p(\theta)$$

and use a MH step with the full likelihood.

For our simulation exercise, and to facilitate comparison with the full GP results, we decided to use the same priors, namely:

Parameter	Priors	Mean	Std Deviation
β	Beta(9.9, 1.7)	0.85	0.099
δ	Beta(1, 10)	0.091	0.083
θ	Beta(5, 10)	0.333	0.117
ρ	Beta(9.9, 1.7)	0.85	0.01
A	G(10, 0.5)	5	1.581
γ	G(0.25, 0.1)	0.025	0.05
l	G(2, 2)	4	2.828
σ_i^2	Inv-G(2.5, 0.5)	0.333	0.471
q_i^2	Inv-G(2.5, 0.5)	0.333	0.471

Table 5.1: Priors for Structural Parameters

In Table 5.1, the parameter l is the length-scale parameter for the Matérn covariance function type, and contrary to Svensson et al. (2016), where l was determined by the user, in our work we allow data to determine the value of l , in a full Bayesian way. Although, after our simulations in the Chapter 4, we now have an inkling of how this prior on l will behave, and could adapt it based on that information, to maintain the comparison between the different chapters' results, we opted to keep the same prior.

For W and Σ , we shall use the priors stated in the previous sections, namely

$$\begin{aligned} W \mid \Sigma &\sim \mathcal{MN}(0, \Sigma, V) \\ \Sigma &\sim \mathcal{IW}(l_\Sigma, \Lambda_\Sigma), \end{aligned}$$

with V as in section 5.1.2, $l_\Sigma = 10$ and $\Lambda_\Sigma = I_3$

Determining the priors and proposal for this method seems to be less demanding, specially since we now have less parameters to worry about.

5.2.3 Proposal Distribution

For the proposal we tried a Metropolis-Adjusted Langevin Algorithm, but similarly to what had occurred before, we observed numerical instabilities in the computation of $\mu(\theta)$ which jeopardized the convergence for the MCMC method. Therefore, we again decided for a Random-Walk type of proposal.

Let us define $\theta = (\theta_{\text{econ}}, \theta_{\text{var}}, l_{GP})$, where θ_{econ} is the vector with the 6 parameters with direct economic interpretation, but this time θ_{var} is the vector with the main diagonal elements of the error covariance matrix Q only, and l_{GP} is the range parameter of the Covariance function of the GP. The proposals for θ_{econ} , θ_{var} and l_{GP} will be again a multivariate and a univariate t distribution:

$$\begin{aligned} q(\theta_{\theta_i} \mid \theta, y, x) &= p_T(\theta_i \mid \mu(\theta_i), s_2 I_{\dim(\theta_i)}) \\ \mu(\theta_i) &= \theta_i + s_1 U, \end{aligned}$$

where $U \sim \text{Unif}(-1, 1)$, and s_1, s_2 are values to be defined by the user, in order to ensure a convergence to the posterior distribution.

Similarly to the previous chapter, with this proposal there is a mismatch between the parameters' space and the proposal's. Instead of just using our priors and proposals as they are defined up to now, obtaining less than desirable rates of acceptance, we decided again to reparametrise the model.

Adapting the reparametrisation from the previous chapter to the current

setting, we now obtain:

$$\begin{aligned} (\tilde{\beta}, \tilde{\delta}, \tilde{\phi}, \tilde{\rho}, \tilde{A}, \tilde{\gamma}, (\tilde{q}_i^2)) &= F(\beta, \delta, \phi, \rho, A, \gamma, l_M, (q_i)) \\ &= (\text{logit}(\beta), \text{logit}(\delta), \text{logit}(\phi), \text{logit}(\rho), \log(A), \log(\gamma), \log(l_M), (\log(q_i^2))) \end{aligned}$$

This will result in the Jacobian

$$\left| \frac{\partial F^{-1}}{\partial \tilde{\theta}} \right| = \frac{e^{\tilde{\beta}}}{(1 + e^{\tilde{\beta}})^2} \frac{e^{\tilde{\delta}}}{(1 + e^{\tilde{\delta}})^2} \frac{e^{\tilde{\phi}}}{(1 + e^{\tilde{\phi}})^2} \frac{e^{\tilde{\rho}}}{(1 + e^{\tilde{\rho}})^2} e^{\tilde{A}} e^{\tilde{\gamma}} e^{\tilde{l}_M} \prod_i^2 e^{\tilde{q}_i^2}$$

Therefore, in the previous modelling, we will think of $\theta = F^{-1}(\tilde{\theta})$ where the only parameters which are changed are those who have just been stated above.

5.3 Results

5.3.1 Preliminary Simulation

In this subsection, we do an analogous exercise to the one done in the previous chapter. We use the data of the last 15 observations in Ireland (2004). Starting from 10 observations, we predicted for 11th period, y_t , and then reestimated the model with 11 observations, and so on.

We used 20 particles in the PGAS, a sample size of 55,000, $s_1 = 0.01$ for the economic parameters MH step, and $s_1 = 0.0001$ for the remaining, and $s_2 = 0.1$ for all MH steps. We used different s_1 step sizes to optimize our results. Similarly, we have also used a burn-in of 5,000.

This time, however, we have two new types of parameters to define, namely (m_1, m_2) which determine the degree of the functional approximation, and (L_1, L_2) which determine the domain on which the approximation will be done. The m_i will have a noticeable effect on the computational time. The better the approximation, the more time consuming its estimation will be. In reference Solin and Särkkä (2014), some values for a good approximation are suggested, at least in the context of GP regression, and thus we decided to follow their suggestion and set $m_1 = m_2 = 12$.

The determination of the parameters L_i is much more complex. Even though we had the results in the previous chapter to guide us, the two models are not exactly the same, since for example our Σ is no longer a diagonal matrix. Several simulations, showed the results to be highly dependent on the considered input space size determined by L_1 and L_2 . In the previous chapter simulations with a full GP for the same data, the estimated state draws would be all between $[-7, 7]$. We tried greater values for L_i such as 20

or 50, and those would degrade the precision of our predictions. Therefore, we decided for $L_1 = L_2 = 10$. For the remaining parameters, since we could not find a guiding intuition for choosing their initial values, we simply decided for $L_\Sigma = 10$ and $\Lambda_\Sigma = I_3$.

In this simulation, at all periods, the initial values we used were the same as the ones used for some previous simulations with a full GP:

$$(\beta, \delta, \theta, \rho, A, \gamma) = (0.7, 0.3, 0.6, 0.7, 2, 0.6), l_{GP} = 15, \text{ and } Q = \begin{bmatrix} 1.5 & 0 \\ 0 & 0.6 \end{bmatrix}$$

The Σ matrix is simply drawn from its prior distribution for the initial value.

It took approximately 53 hours for this simulation to finish, using a server with a CPU of 30 cores, whereas in our previous simulation with a full GP, it took approximately 70 hours, with a tendency for this discrepancy to increase as more data is used. This is an improvement of 24% in the computational time.

Traces

To assess mixing and convergence of our chains, we now look at the traces. In the traces below, one may see that there is still leeway for improvement, but the draws obtained are still usable.

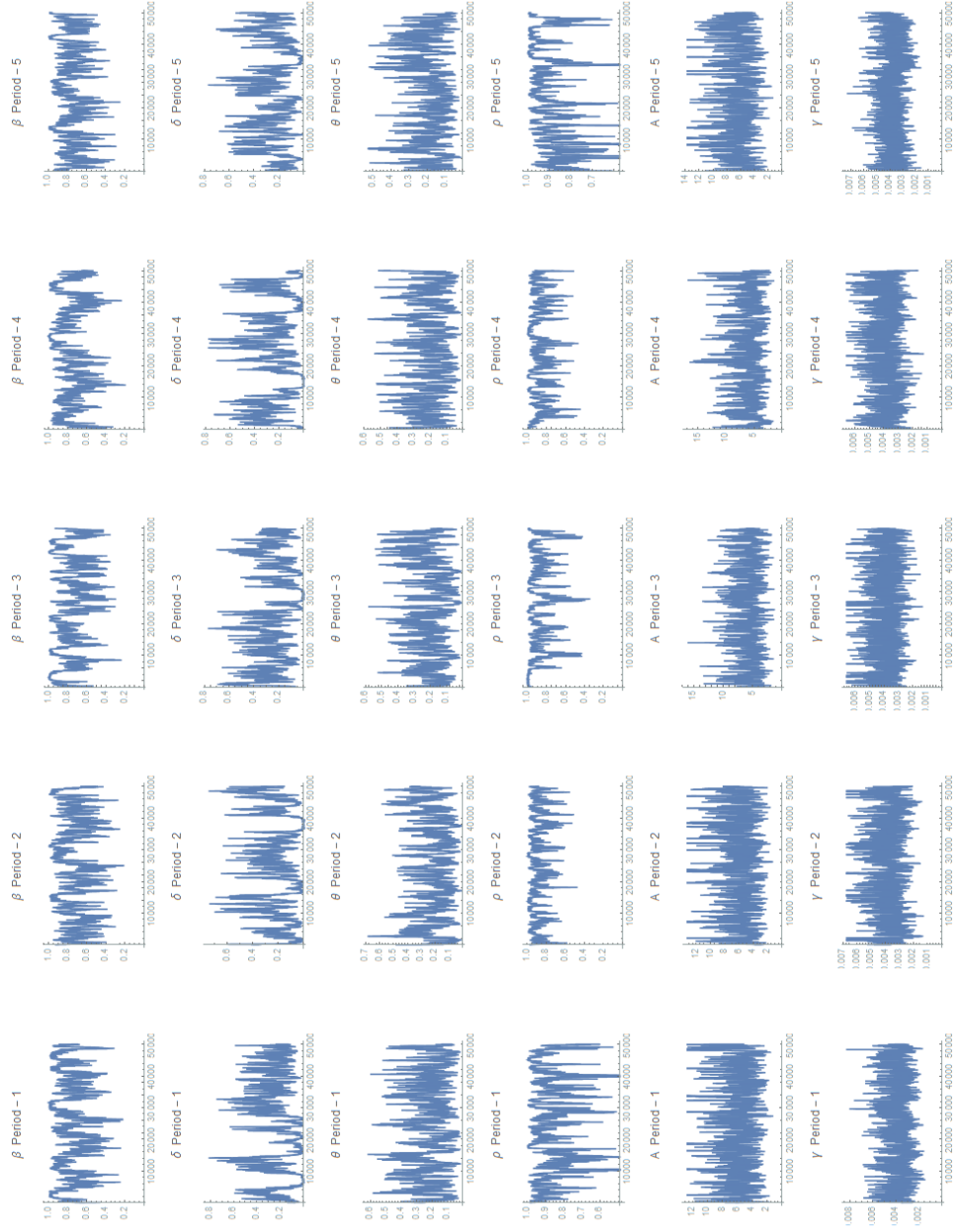


Figure 5.1: Traces of Economic Parameters Draws

The traces for this model seem to suggest more difficulties in mixing, than the full GP, since we can see some chains 'stuck' on certain areas of the parameter space. This is most noticeable in the β , δ and ρ parameters. We tried something akin to gradient descent, however problematic precision issues were recurrent, invalidate our attempt. One important future research component would be how to improve the mixing of these parameters.

Next, we shall analyse the trace for the elements q_1 and q_2 .

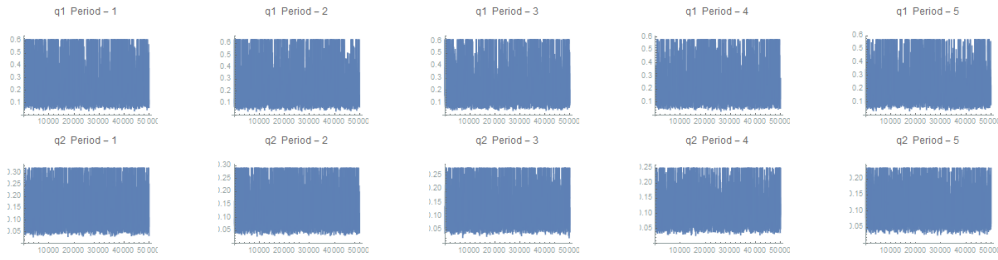


Figure 5.2: Traces of q_1, q_2 Parameters Draws

They mix very well. Contrasting with a previous analysis with a full GP, one should notice that this current approximative method uses a conjugate prior, so for each iteration of the algorithm we are certain to get a new realization Σ . Furthermore, we are using a more complex matrix, instead of a simple diagonal Σ as it was previously assumed.

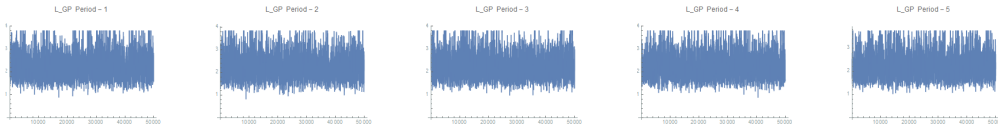


Figure 5.3: Traces of l_{GP} Parameters Draws

Another difference from our previous full GP estimation, the l_{GP} trace seems to have explored smaller values. We will confirm this in the histograms below. One possible cause could be that now the Σ is no longer diagonal, and incorporates more non-null terms. Hence, the error term is more flexible, and may account for some features of the data that were not accountable in the previous analysis.

Posterior Histograms, and Density Priors

Similarly to what was previously done, we will now analyse the histograms from our estimated posterior distributions for each parameter.

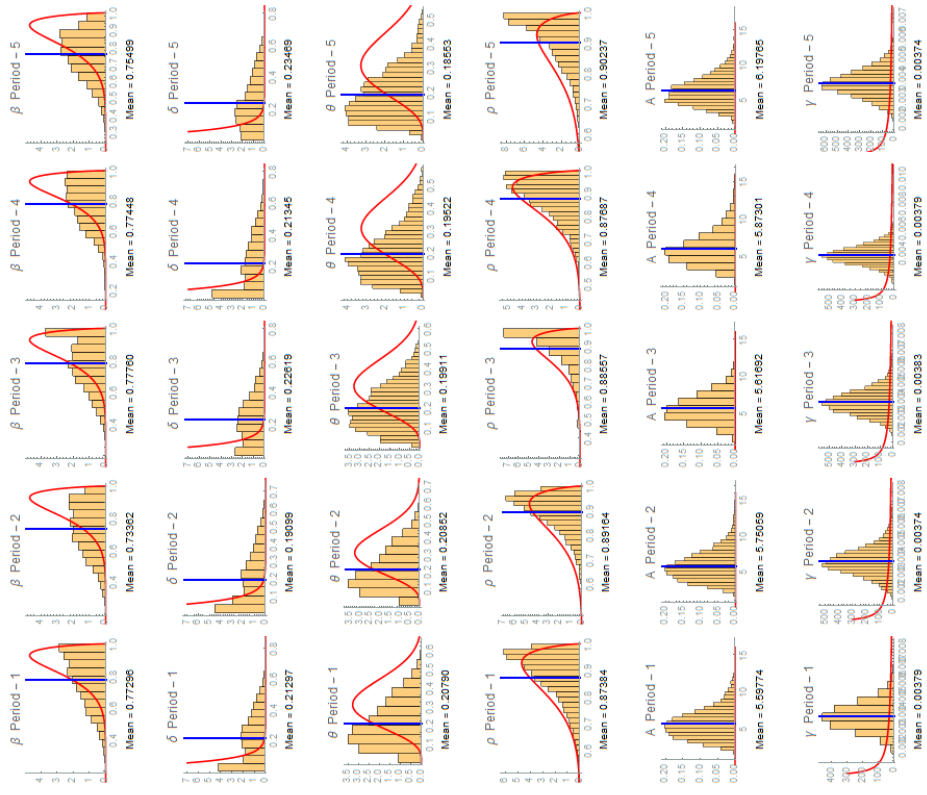


Figure 5.4: Histograms and prior densities for structural parameters

From Figure 5.4, the data seems to be informative for all parameters, despite the small size of data used. We can also notice some of the difficulties that chain had in mixing, with some histogram shapes being a bit dissimilar as we add just a few data. For example, the histogram for δ parameter at period 3 exhibits a shape which is dissimilar from the previous or the following histogram. The parameter estimated mean value also seems to be in line with the values from the full GP simulations.

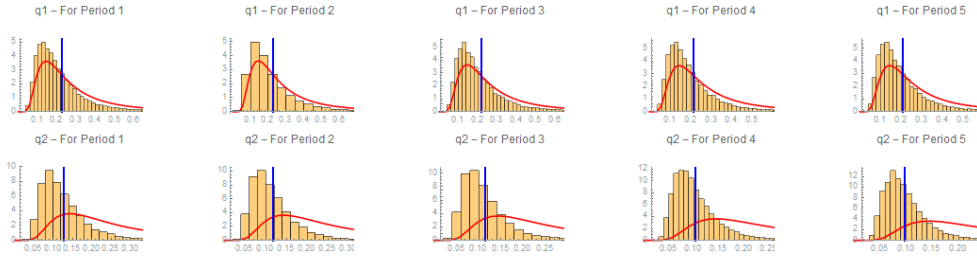


Figure 5.5: Histograms and prior densities

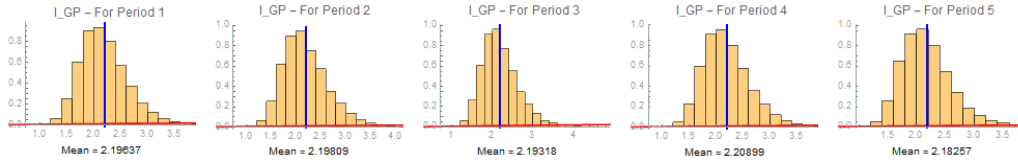


Figure 5.6: Histograms and prior densities

The histograms of Figures 5.5 and 5.6, show our model features also learning from data.

Now we can confirm the posterior distribution for l_{gp} has translated considerably from our previous results with a full GP, by moving from an average in $[5.0, 5.2]$ to an average of approximately of 2.2. This might be a consequence of the values fixed for L_i and m_i .

Predictions for 5 Periods

As one would expect, the HRGP has a RMSE for each of component of y of

$$(0.0368053, 0.0264761, 0.0159962)$$

which is higher than the full GP RMSE, but still considerably lower than y_M RMSE.

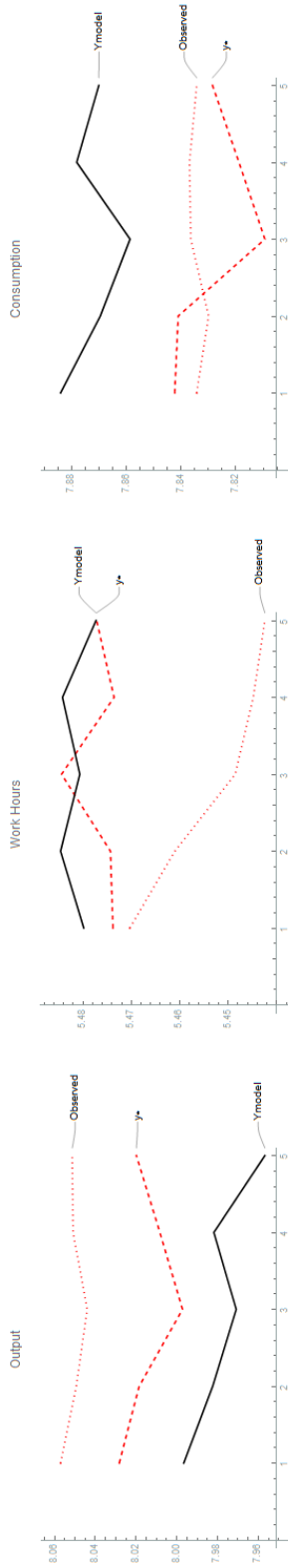


Figure 5.7: Histograms and prior densities

Thus, similarly to our conclusion in the full GP analysis, the HRGP behaves better for consumption, which in turn will be reflected in output, but less so. The RMSE for Work Hours is surprisingly better than the full GP, but still at the same magnitude as the y_M predictions.

5.3.2 Comparing Performances

Here we do a similar exercise to what was done in Ireland (2004), but due to our computational constraints we had to limit ourselves to 1 lag predictions, considering only the previous 10 observations, i.e., for each of the next 69 periods, we used the previous 10 observations to predict the next period.

To make this exercise faster, we only used a sample size of 15,000, with a size burnin of 5,000. The step sizes were similar to those before, $s_1 = 0.0001$, $s_2 = 0.1$, $s_{2\text{Econ}} = 0.01$, and we drew 50,000 times from the posterior predictive distribution. We still used 20 particles, but now $(m_1, m_2) = (8, 8)$ and $(L_1, L_2) = (7, 7)$. We decided to decrease the degree of the approximation to make it faster. The total computation time for this exercise was of about 7 days and 2h. This resulted in an approximate 40h saving regarding a similar exercise done with the full GP.

The resulting RMSE for this exercise was (0.0677465, 0.0288514, 0.0318878). Even though we expected a decrease in predictive performance when compared with the full GP, it was still much better than the results reported by Ireland (2004). This results maybe highly dependent on the choice of the m_i and L_i .

5.4 Conclusions

The main objective of this chapter was to increase the computational tractability of the our UQ of model discrepancy. The method has indeed decreased the time, e.g. a 20% decrease when we ran the simulation exercise for 79 observations of Ireland (2004). Even though as more observations are used, the improvement in computational performance increases, their absolute performance still seems to be prohibitively high. The main improvement in efficiency happens at the sampling of the state distribution, however, since we also added an extra step for the sampling of the W and Σ matrices, and there was decline in the mixing of the chains from economic parameters, a more accentuated efficiency improvement was hindered.

We were expecting a worsening of the predictive performance, and we obtained between 1.5 and 4 times worse RMSE. The HRRGP method is very influenced by the 'calibration' of parameters such as m_i and L_i . Some simulations, but which were not integrated into our work, seemed to indicate a very sensitive dependence on L_i (too big or too low would increase even more the RMSE) and suggest a discrimination for the values associated with the different components of x_t .

A venue of future research would be to improve the proposals of the MH

step when learning the system, and the use of a faster programming language.

The use of online⁸ algorithms, such as those using the variational principle, which itself already uses approximations to the likelihood, accumulated with the approximation of the HRRGP method, risks worsening the forecasting performance of our method, or worse, defeat our purpose of quantifying model misspecification by adding even more sources of uncertainty.

⁸Online algorithms is a Machine Learning designation for algorithms which run in real-time, and hence they usually do not do batch estimation, but as more data is made available, they update the learning using computations done in previous estimations in an optimised way, sometimes using some approximations.

Chapter 6

UQ during the Great Crisis/Recession

6.1 The Great Crisis: A Summary

The official arbiter of U.S. recessions, the U.S. National Bureau of Economic Research, states that the recession began in December 2007 and ended in June 2009. During this period, and according to the International Monetary Fund the «downturn represents by far the deepest global recession since the Great Depression»¹ Some authors also use the term "The Great Recession", however, recession in the sense of lower economic activity, or of a mild slump in the business activity ² does not convey the extent to which this period marked the global economy, ushering in a new world of low interest rates, and diminished growth for developed countries.

The U.S. Financial Crisis Inquiry Commission, on a January 2011 report³, concluded:

- The Crisis was avoidable
- «[The] widespread failures in financial regulation and supervision proved devastating to the stability of the nation's financial markets.»
- «The dramatic failures of corporate governance and risk management at many systemically important financial institutions were a key cause of this crisis.»

¹See International Monetary Fund (2009)

²There is also a more formal definition of recession which is the consecutive decrease of GDP for two or more quarters.

³See The Financial Crisis Inquiry Commission (2011)

- «A combination of excessive borrowing, risky investments, and lack of transparency put the financial system on a collision course with crisis.»
- «The government was ill prepared for the crisis, and its inconsistent response added to the uncertainty and panic in the financial markets.»

In the years previous to the onset of the crisis, the U.S. economy was marked by a bubble in asset prices driven primarily by an inflow of savings from developed nations, which was funnelled into the mortgage market.

Until recently, it was thought most of the market had been driven by subprime lendings to poor and low credit rating households. In Adelino, Schoar, and Severino (2016), a distinct characterization of that market is put forth. In this article, the authors show how the high and middle-income borrowers comprised the majority of the mortgage market, and during the Great Crisis, these households accounted for a disproportioned share of the delinquent/non-performing loans. The reason for this is the greater magnitude of the average mortgage to middle to high-income households with respect to those of low-income households. So, a small increase in their default rate had a severe impact in the mortgage market. Until 2006, the U.S. mortgage-backed securities were very desirable assets due to yields being higher than the U.S. government bonds, while being considered fairly safe, and very liquid (traded globally). In 2007, mortgage owners started to default in great numbers, making heavy weight investors in this market to accumulate losses. Adding to this, the U.S. shadow banking system, namely investment banking and other non-depositary financial institutions, were not subject to a regulatory framework such as regular banks, which made them vulnerable to market panic and consequent bank runs. The event triggering this last scenario in the shadow banking system was Lehman Brothers filing for bankruptcy in 2008. The dependencies among these investors facilitated contagion across systems (commercial banks, insurers such as AIG, etc) and across countries. When European governments were called to bail-out commercial banks, their public debt surged which created the 'Sovereign Debt Crisis'.

6.2 DSGE Models during the Great Crisis

It is known that DSGE model based forecasts failed to predict ex-ante⁴ the great recession, whether in its magnitude, whether in its duration, besides

⁴Previous to the event, given the data at that time

the references already stated in the introduction to this work. See also, for example Wieland et al. (2012).

The DSGE models usage has several objectives, such as policy decision making tool, counter-factual experiments, or supplying an economic interpretation of facts. This role is usually not possible with the usual statistics only based models. However, it was only relatively recently that interest in the forecasting performance of DSGE models surged, specially for the period of the great crisis.

One such paper is Negro and Schorfheide (2013), where the authors test three models: the Smets and Wouters model of Smets and Wouters (2003) and Smets and Wouters (2007), this SW model with Financial Frictions such as those seen in Bernanke, Gertler, and Gilchrist (1999), and a small-scale DSGE model which is a simplification of the SW model. The authors test these models against the forecasts published in the Federal Reserve Board of Governors("Greenbook") and professional forecasts published in the Blue Chip survey(BCS). Even if we were to integrate information from the nowcasts⁵ of the BCS in the DSGE forecasting, these latter models still perform worse when compared with professional forecasts in the short-run range. Only at medium and long-run do they perform competitively with the BCS forecasts, if we also incorporate long-run inflation expectations. Although not stated in their work, this comparison of a non-micro-founded altered DSGE model, seems an implicit assumption that DSGE models are clearly lacking in reality checking, similarly to the conclusions in the work of the same authors with DSGE-VAR forecasts. When proceeding to analyse the great crisis period - using the SW model with long-run expectations(LRE), SW+LRE+financial frictions, the SW+LRE+financial frictions+nowcast, they conclude that in general all models failed to predict the magnitude(4th quarter of 2008) and extension of the crisis. The notably exception was the SW+financial frictions+nowcast model for the 90% bands for output growth **only**. The 80% band missed the target, as well as any DSGE forecast band for inflation. This exercise also allowed them to rank the models, with SW+LRE+financial frictions+nowcast coming out as the preferred model to forecast during a the great crisis.

However, this forecasting performance is not stable, since outside the crisis period, the SW+LRE gave more accurate predictions. This could be why these authors gave an implicit indication of the use of Bayesian Model

⁵In Macroeconomics, some time series take a long time to be obtained, and even then they may be subject to subsequent revisions. Nowcasting refers to the prediction of the near future, present, or even the very recent past, for which such time series are still not available, but many other indicators are. It can be thought of real-time tracking and forecasting.

Averaging in some of the references stated in Chapter 1 of our work.

6.3 Can our UQ framework capture the Crisis?

Similarly to Ireland (2004), the majority of our data will be taken from Federal Reserve Bank of St. Louis' FRED database. We will use for Y_t the real gross domestic product, billions of chained 2012 dollars, quarterly, seasonally adjusted annual rate (GDPC1). For C_t , we used real personal consumption expenditures, billions of chained 2012 dollars, quarterly, seasonally adjusted annual rate (PCECC96). For H_t , we decided to use hours of wage and salary workers on nonfarm payrolls: private sector, billions of hours, quarterly, seasonally adjusted annual rate (PRSCQ). All series are converted to per-capita terms by dividing by the civilian, non-institutional population level, age 16 and over, by thousands of persons, monthly (we only use the values at months 1,4,7 and 10), not seasonally adjusted. We use data from the period of 1964Q1 to 2010Q1.

Contrary to the Ireland (2004) method, ours does not detrend the data automatically, and in many of the previous exercises we simply assumed a value of 1.0051 for η . For this chapter, as in the last simulation done in Chapter 4, we will proceed differently, estimating η from all the past data available to us until the current estimation period. We estimate η up to period T , then in the next estimation period, we use up to period $T+1$ data. For the reader's convenience, we repeat succinctly how to proceed with the detrending.

To estimate this parameter we use the definition of $y_t = \frac{Y_t}{\eta^t}$ and so, we are expecting the following relationship:

$$\log(Y_t) = t \log(\eta) + \log(y_t)$$

So, we regress all previous log output data on t in a linear model of the form

$$\log(Y_t) = \beta_0 + t\beta_1 + \epsilon_t$$

Using this technique, for data up to 2007Q4, we got $\exp \beta_1 = 1.00429$ which is a reasonable value as a sanity check. Now, we estimate the model for 2007Q4, we do a one-step forecast, and then reestimate η with data up to 2008Q1, and so forth.

6.3.1 Data Plots

For a general idea of the data we will be working in this chapter, see Figure 6.1.

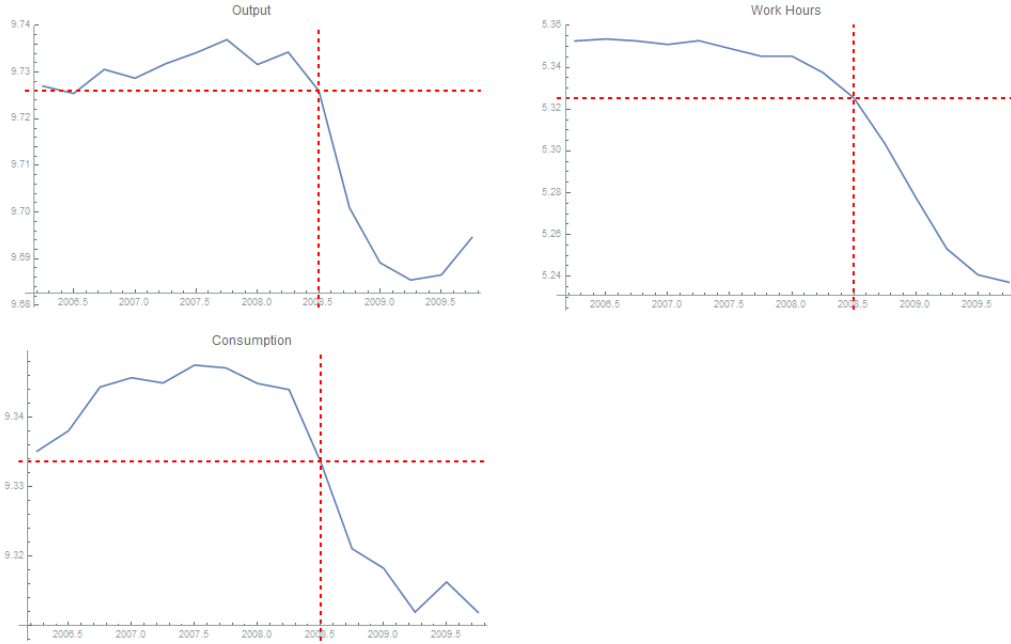


Figure 6.1: Plots for different components of \hat{m}

To the leftmost part of each red cross, is the data which will be used to predict the rightmost part of each red cross. It bears some resemblance to the simulations we conducted in previous chapters but with a greater downturn. The reason for this is that in the Ireland (2004) dataset, the last 15 observations corresponded to the 2001 dot com bubble burst, and now we are forecasting 2008Q3 to 2009Q3, comprehending the one of biggest turmoil of economic fundamentals in recent history, the great world crisis⁶. With this graph we can already foresee that our forecasting exercise will be more akin to extrapolation than interpolation, and so we can expect the method to decrease its performance with respect to the previous exercises.

⁶At the time of this writing, the Coronavirus Pandemic is still in its infancy. The exact consequences of this never before seen event in human history, will change the world, and can only be assumed/guessed at this point. Hence, we will disregard it by now, due to its uncertain outcome

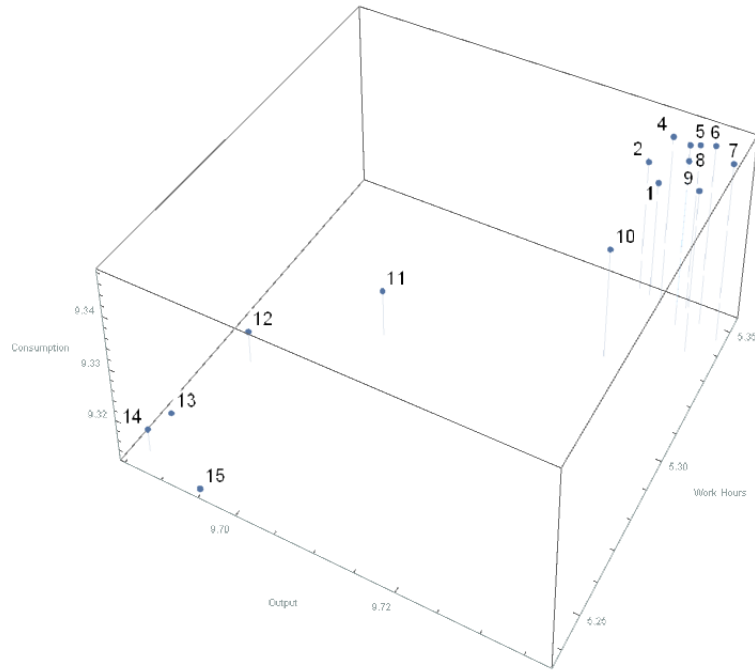


Figure 6.2: A 3D Plot of Observations: Before and During The Great Crisis

The 3D plot makes the difficulty of this forecasting exercise very stark. By observing Figure 6.2, one intuitively sees just how problematic this forecasting exercise may become. The forecasts for the 11th, 12th, and 15th data points seem to be potentially the most problematic ones, since those are the farthest from previous data points.

6.3.2 Traces

Proceeding in this manner, with the same exercise input program values setting of the Full GP simulations as in Chapter 4, we obtain the following traces.

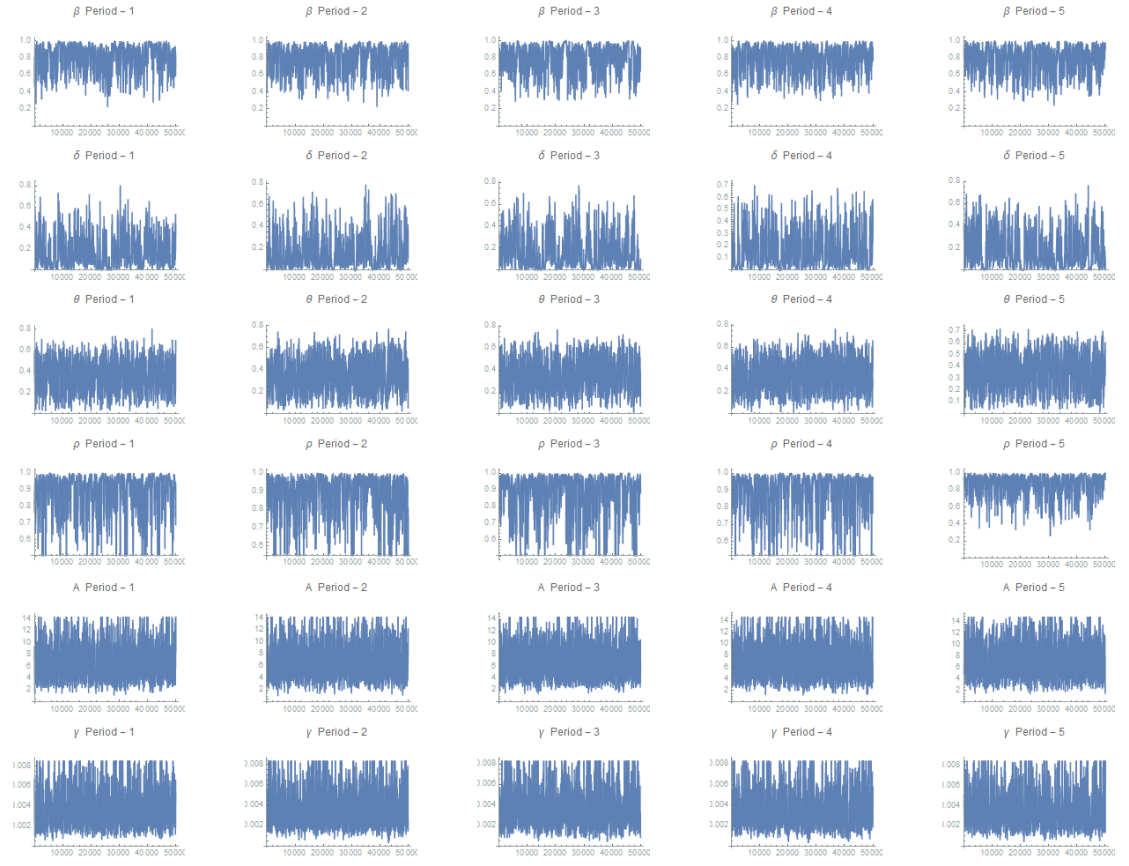


Figure 6.3: Trace of Economic Parameters

In Figure 6.4, we have the traces for each element of M_1 . The mixing seems better than in the previous full GP simulations, in Chapter 4, but still with some room for improvement.

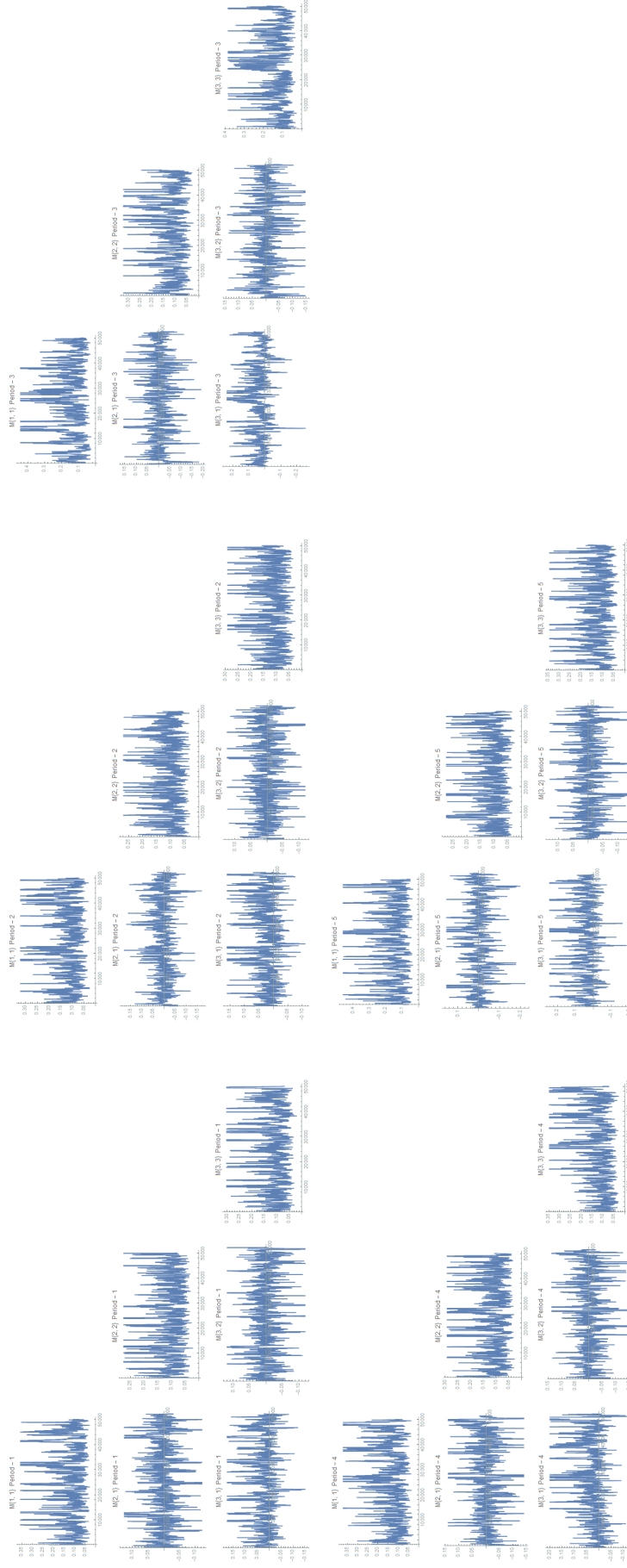


Figure 6.4: Trace of M_1 componentwise

Figures 6.5 and 6.6 show the traces for the remaining parameters. The mixing seems to be good as can be seen in the below figures.



Figure 6.5: Trace of σ_i^2, q_i^2

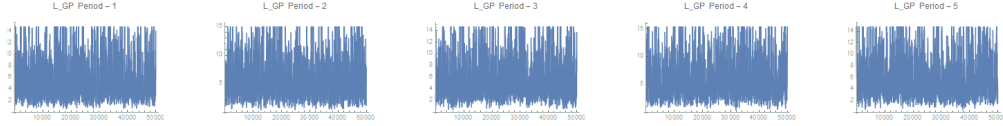


Figure 6.6: Trace of l of the Covariance Function

6.3.3 Posterior Histograms and Density Priors

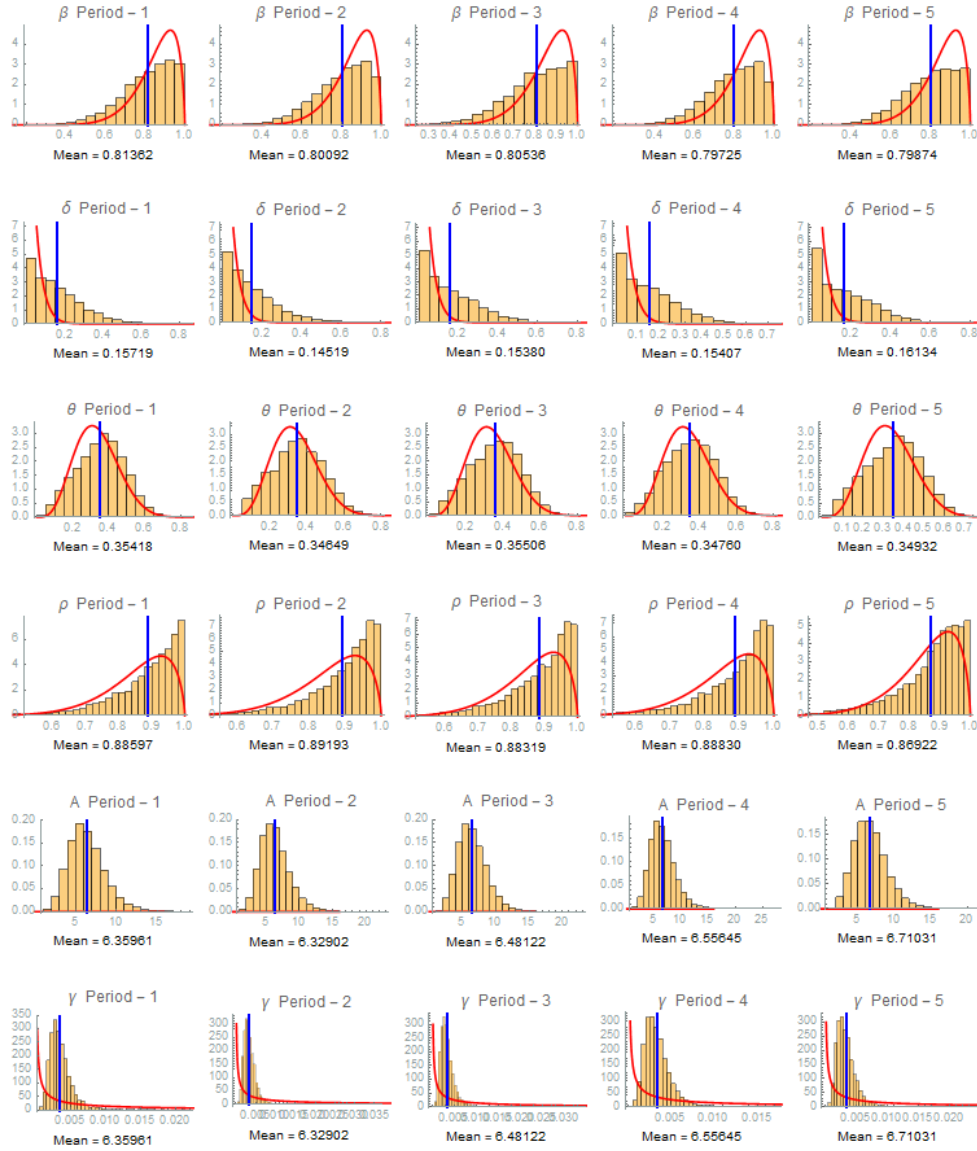


Figure 6.7: Histograms and Density Priors for Economic Structural Parameters

The histograms in Figure 6.7, now confirm mixing is quite good. Similarly to previous exercises, the inclusion of a discrepancy term in the observation equation with a GP prior is still not enough to obtain parameter estimations more in line with scientific economic knowledge. For example, the β coefficient is expected to be higher than 0.9.

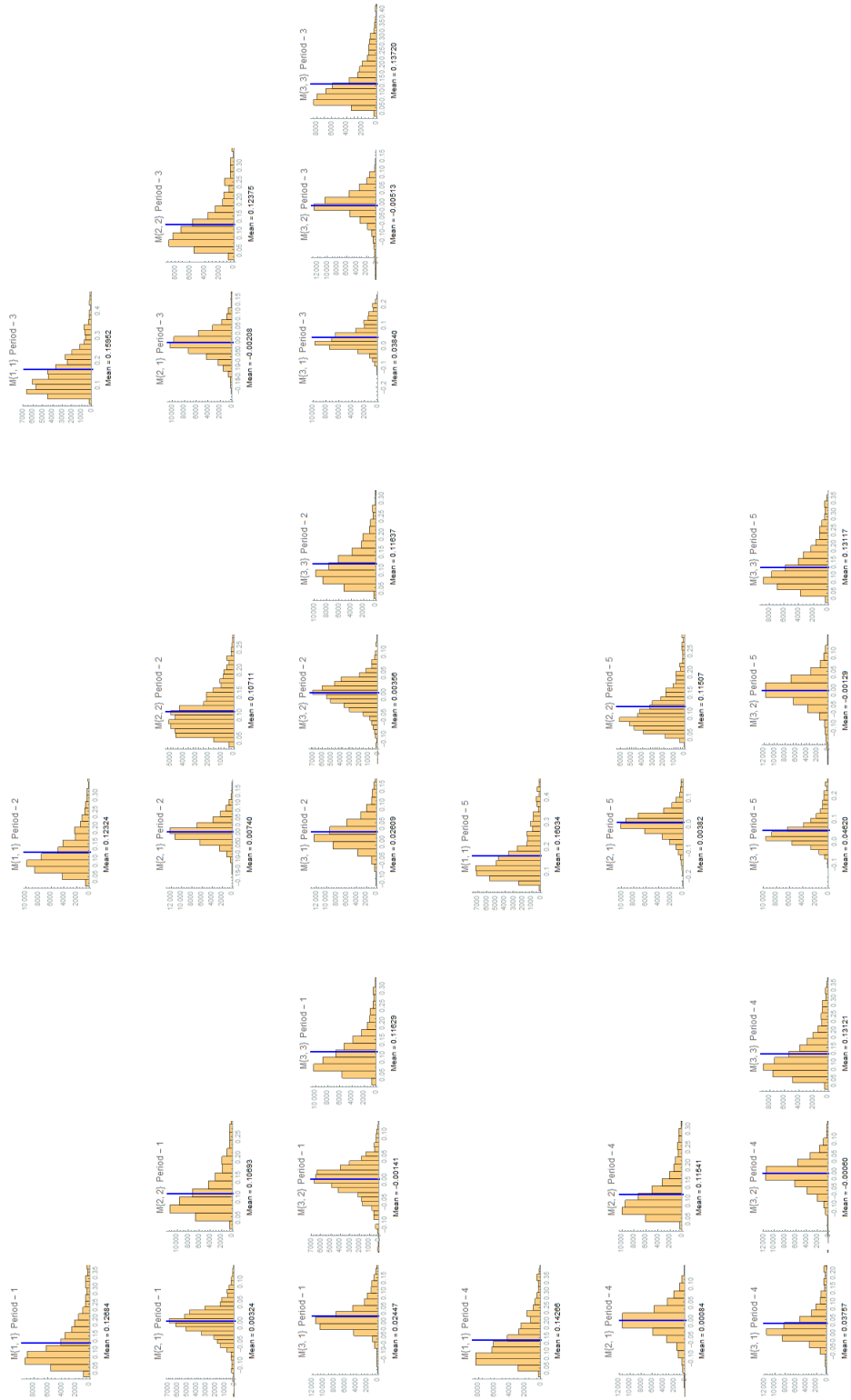


Figure 6.8: Histogram of M_1 componentwise

Period 1	Period 2	Period 3
$\begin{pmatrix} 1 & & \\ 0.0278 & 1 & \\ 0.2015 & -0.0126 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & & \\ 0.0644 & 1 & \\ 0.2178 & 0.0319 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & & \\ -0.0148 & 1 & \\ 0.2596 & -0.0394 & 1 \end{pmatrix}$
Period 4	Period 5	
$\begin{pmatrix} 1 & & \\ 0.0066 & 1 & \\ 0.2746 & -0.0049 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & & \\ 0.0281 & 1 & \\ 0.3186 & -0.0105 & 1 \end{pmatrix}$	

Table 6.1: Correlation matrices

In this exercise the same relationships between components exist, i.e. Consumption and Output are still very correlated, more so than in Chapter 5.⁷, and Work Hours is uncorrelated with any component.

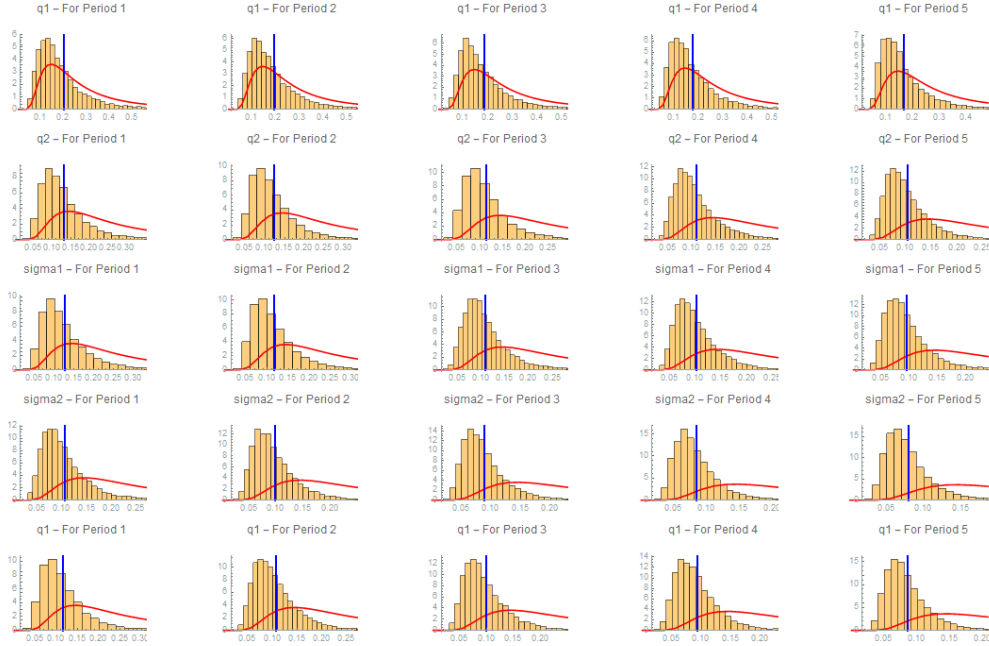


Figure 6.9: Histograms and Density Priors for σ_i^2, q_i^2

⁷Notice that estimation periods 2 to 5 correspond to the forecasting periods 1 to 4. So, the last observation, in forecasting period 5, where Output and Consumption differ markedly is not present here.

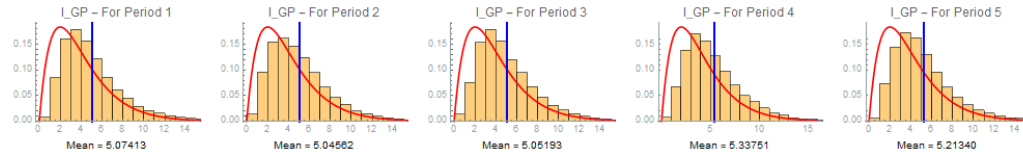


Figure 6.10: Histograms and Density Prior for l of the Covariance Function

The histogram for the l_{GP} seems very similar to that in Chapter 4, even the mean is very similar.

6.3.4 Forecasts

1-step forecasts

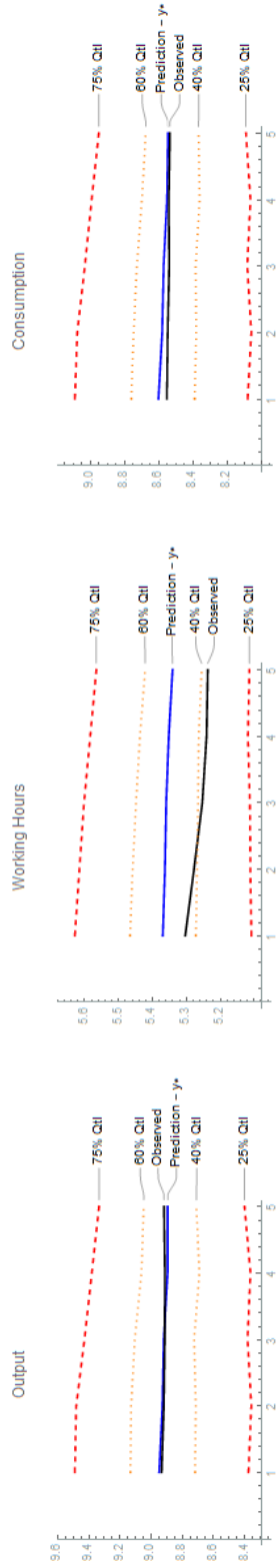


Figure 6.11: Predictive plot for 5 periods

In Figure 6.11, we can see several quantiles computed from the posterior predictive distribution for each component. These 40% to 60% quantile form a prediction interval containing 20% of the mass of the posterior predictive distribution. The 25% to the 75% quantiles form a prediction interval 50% of the mass of the posterior predictive distribution. The 20% prediction intervals capture almost all of the variation except for the Work Hours component from the 3rd period onwards. This behaviour of assigning a very large uncertainty may be the result of our chosen covariance function, but it may also be the due our reduced sample size, and/or a relatively vague prior. In the next subsection, we will investigate this issue further.

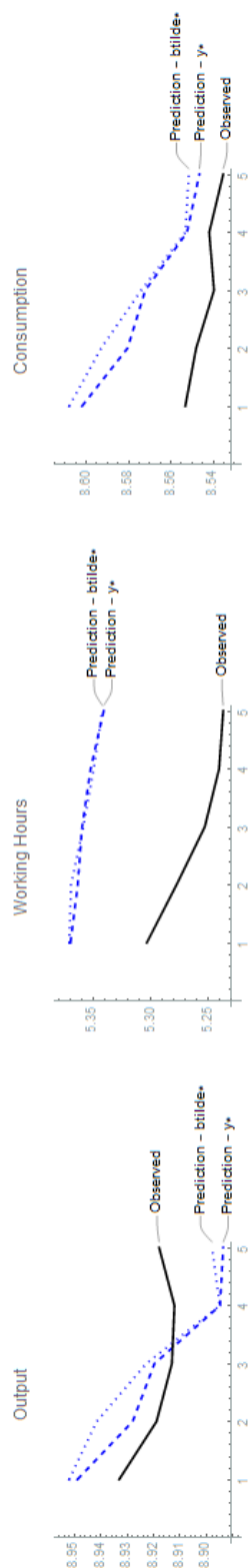


Figure 6.12: Two different ways of prediction

In Figure 6.12 we can clearly see that Output and Consumption forecasts are significantly correlated.

The RMSE of our 1-step forecast is (0.0159038, 0.0962564, 0.030751) with the biggest error associated to the work hours, just like in the simulations with the full GP in Chapter 4.

6.4 Some Tighter Priors, and a Smoother Covariance Function

In this section, we will do a similar forecast exercise, but with tighter priors on some parameters.

Parameter	Priors	Mean	Std Deviation
β	Beta(10, 0.7)	0.935	0.072
δ	Beta(0.5, 25)	0.0196	0.027
θ	Beta(11, 22)	0.333	0.081
ρ	Beta(10, 0.7)	0.935	0.072
A	G(16, 0.3)	4.8	1.2
γ	G(0.25, 0.1)	0.025	0.05
l	G(1.8, 3)	5.4	4.025
σ_i^2	Inv-G(2.5, 0.5)	0.333	0.471
q_i^2	Inv-G(2.5, 0.5)	0.333	0.471

Table 6.2: New Tighter Priors

Some of the main differences were in the β and ρ priors.

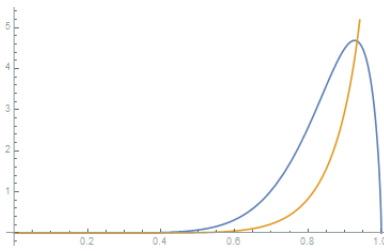


Figure 6.13: Original Prior and New tighter Prior for β and ρ

In Figure 6.13, in blue we see the original prior for β and ρ , and in orange we see the new tighter prior for both of those parameters. We decided for a density that would have a less pronounced left-side tail.

Another great change was in the prior of the length-scale parameter l_{GP} of the covariance function. In this subsection, our first simulation uses a

$G(1.8, 3)$ distribution which has a smaller mean mean but a much bigger standard deviation, approximately twice as much.

For the matrix M_1 , we decided to use $\text{Inv-Wish}(I_3, 15)$. All the priors, except for the one on l and M_1 , could be considered tighter than before. For the exercise in this section, we decided to use two different Covariance Functions. Our first one is the Matèrn covariance function using $\nu = 5/2$, making the process at least 2 times M.S. differentiable, meanwhile keeping the simplicity of the covariance function:

$$k_{\nu=5/2}(r) = \left(1 + \frac{\sqrt{5}r}{l} + \frac{5r^2}{3l^2}\right) \exp\left(-\frac{\sqrt{5}r}{l}\right)$$

Our second one is the Squared Exponential covariance function, making the process infinitely M.S. differentiable:

$$k(r) = e^{-\frac{r^2}{2l^2}}$$

One expects that a smoother covariance function will impose a greater tightening of the prediction intervals.

6.4.1 Results For Matèrn Smooth Covariance Function

To avoid cluttering this section with graphics, it will suffice to say that the mixing behaviour of our chains with the new priors were very similar to previous ones with the full GP.

The main differences were with respect to the posterior distributions of some parameters.

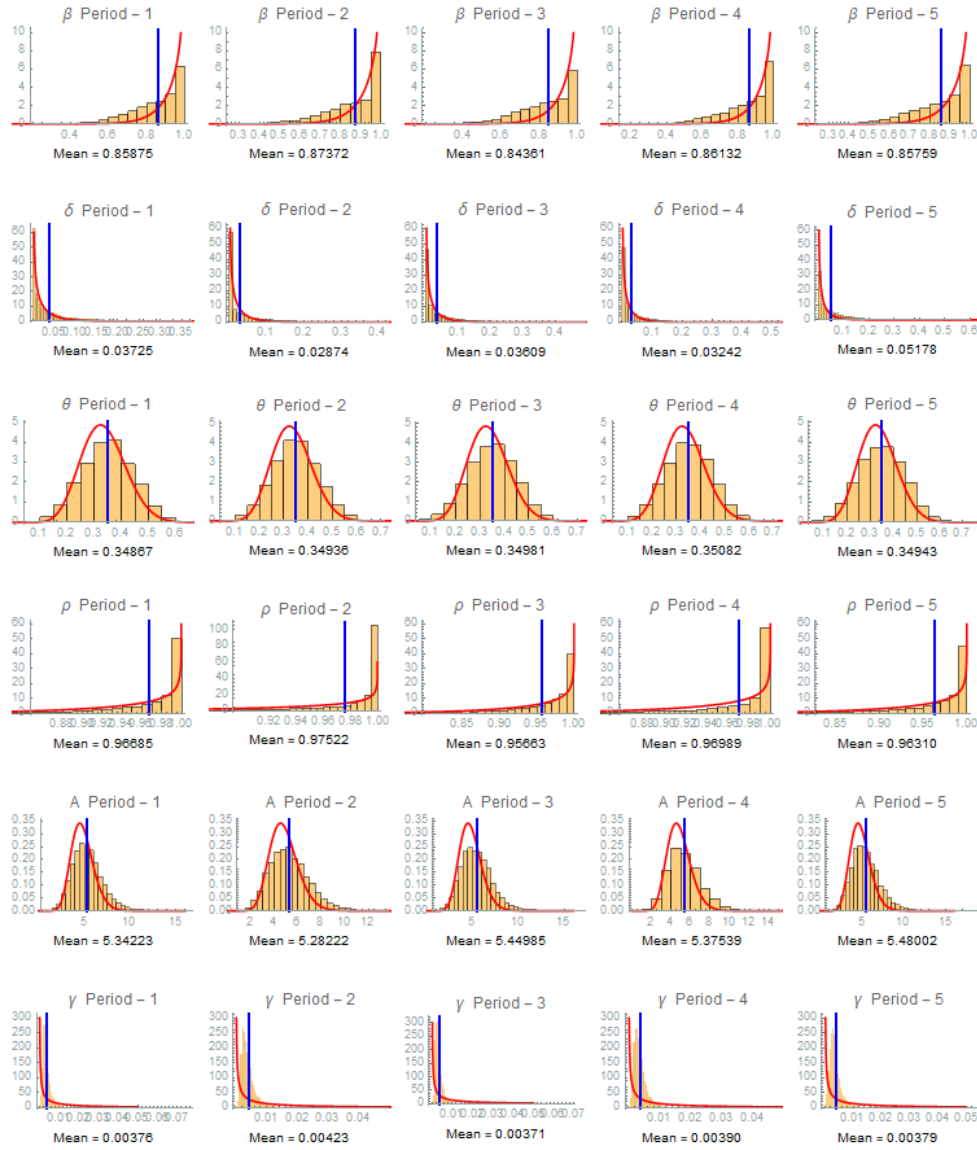


Figure 6.14: Posterior Histograms and Density Priors for Economic Parameters

Generally, since we were dealing with few observations, the priors have imposed themselves on the posterior distribution in a much stronger way, removing some of the data influence we had in previous simulations. Nevertheless, we can still glean some information from data. For example, regarding parameter A and γ , as we add more data, an increasing difference between the prior and estimated posterior shows itself. Furthermore, ρ 's posterior distribution, which we saw had a tendency to position itself closer to 1, with

the change of prior, moved even more pronouncedly to the upper bound of its interval.

Another main difference in this exercise is how the different components of the GP are related amongst themselves.

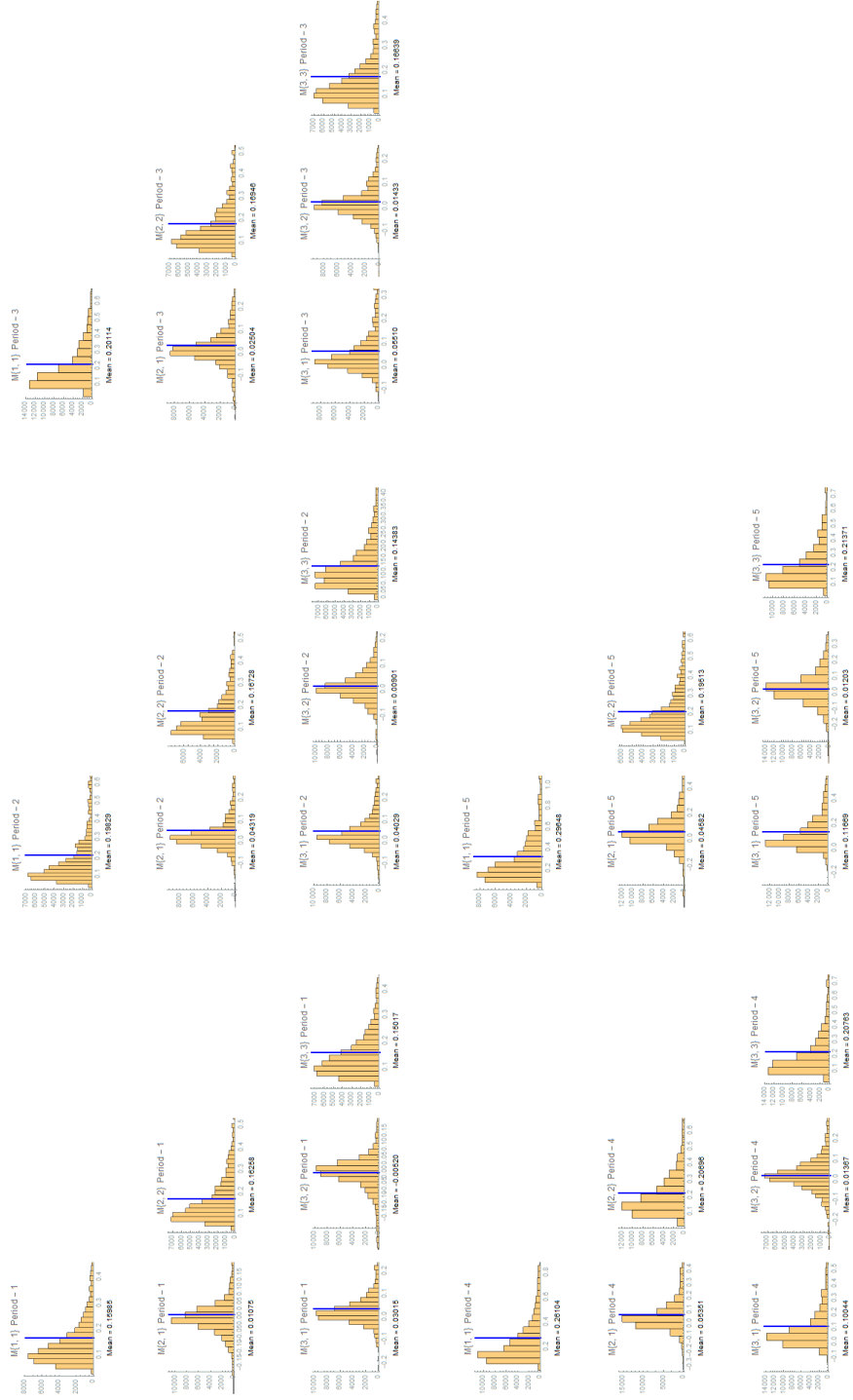


Figure 6.15: Histogram of M_1 componentwise

We can see from the Table 6.3 a surprising result. Notice that component 2(Work Hours) now has some non-negligible correlated behaviour with component 1(Output).

Period 1	Period 2	Period 3
$\begin{pmatrix} 1 & & \\ 0.0667 & 1 & \\ 0.1946 & -0.0333 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & & \\ 0.2372 & 1 & \\ 0.2386 & 0.0581 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & & \\ 0.1356 & 1 & \\ 0.3012 & 0.0853 & 1 \end{pmatrix}$
Period 4	Period 5	
$\begin{pmatrix} 1 & & \\ 0.2302 & 1 & \\ 0.2431 & 0.0659 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & & \\ 0.1905 & 1 & \\ 0.4636 & 0.0589 & 1 \end{pmatrix}$	

Table 6.3: Correlation Matrices for Smoother Covariance Function

If we compare the posterior distribution for l_{gp} we will see that with the new covariance function, the posterior distribution adapts faster to the inclusion of new data.

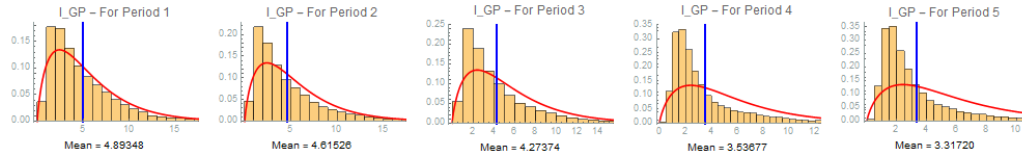


Figure 6.16: Histograms and Density Prior for l of the Covariance Function

The mean of the posterior distribution of l_{gp} is more reactive as we add more data. It tends toward smaller values than when we used a rougher covariance function. A smaller value of the length-scale makes the process less smooth. Hence, the increased tendency of smaller values of l_{gp} may be a way to compensate for our newly imposed increased smoothness for the full GP.

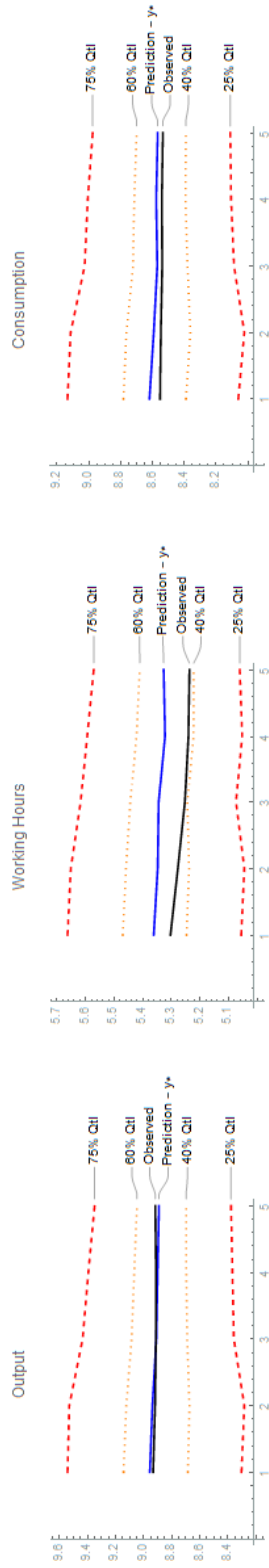


Figure 6.17: Posterior Predictive Intervals for y^*

In Figure 6.17 we see that the imposition of our new covariance function was not enough to decrease by much the predictive interval magnitude, at least initially. There is a clear tendency for a greater tightening as more data is added, than we previously had, though. In Figure 6.22, we will clearly see how the length of the predictive intervals behaves according to the smoothness of the covariance function, and across periods.

The resulting RMSE, (0.0188905, 0.079492, 0.0440202) reflects the increased correlation of the Work Hours with Output, making this RMSE associated with the former smaller than previous 0.0962564. The remaining RMSE errors had a slight increase, probably due to the vagueness of the length-scale prior.

6.4.2 Results For Squared Exponential Covariance Function

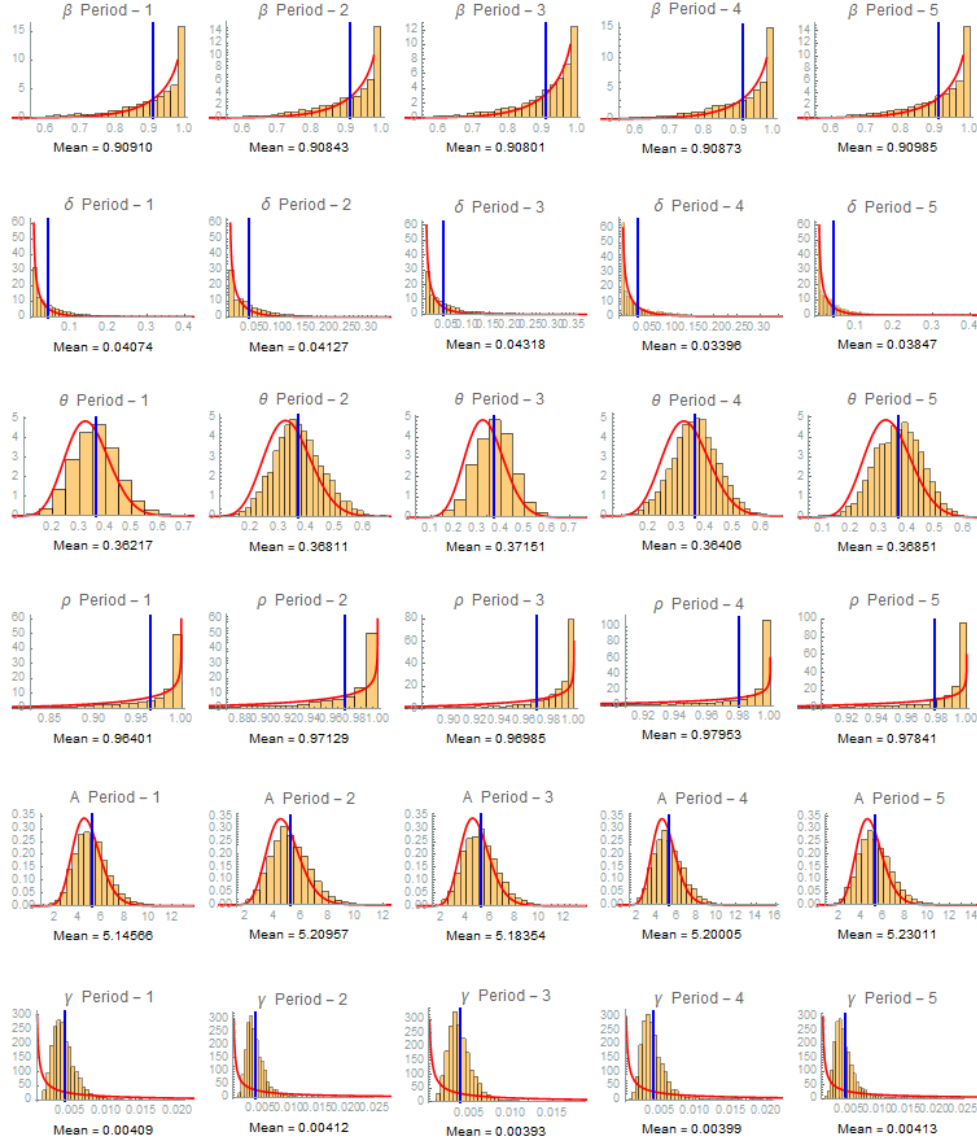


Figure 6.18: Posterior Histograms and Prior Densities for Economic Parameters Using Infinitely Smooth Covariance Function

Similarly to Matérn Smooth Covariance function, and since we were dealing with few observations, the priors imposed themselves on the posterior distribution in a strong way. Nevertheless, the posteriors have become tighter, than in any of the previous simulations with the Matérn Covariance, as can

be seen for example in the posterior on ρ or β . Hence, we can expect the posterior predictive distribution to also be tighter than in the previous simulations.

With the Matèrn smooth simulation we saw Work Hours to increase its correlation with respect to Output. The next histogram seems to indicate that behaviour is now even more pronounced.

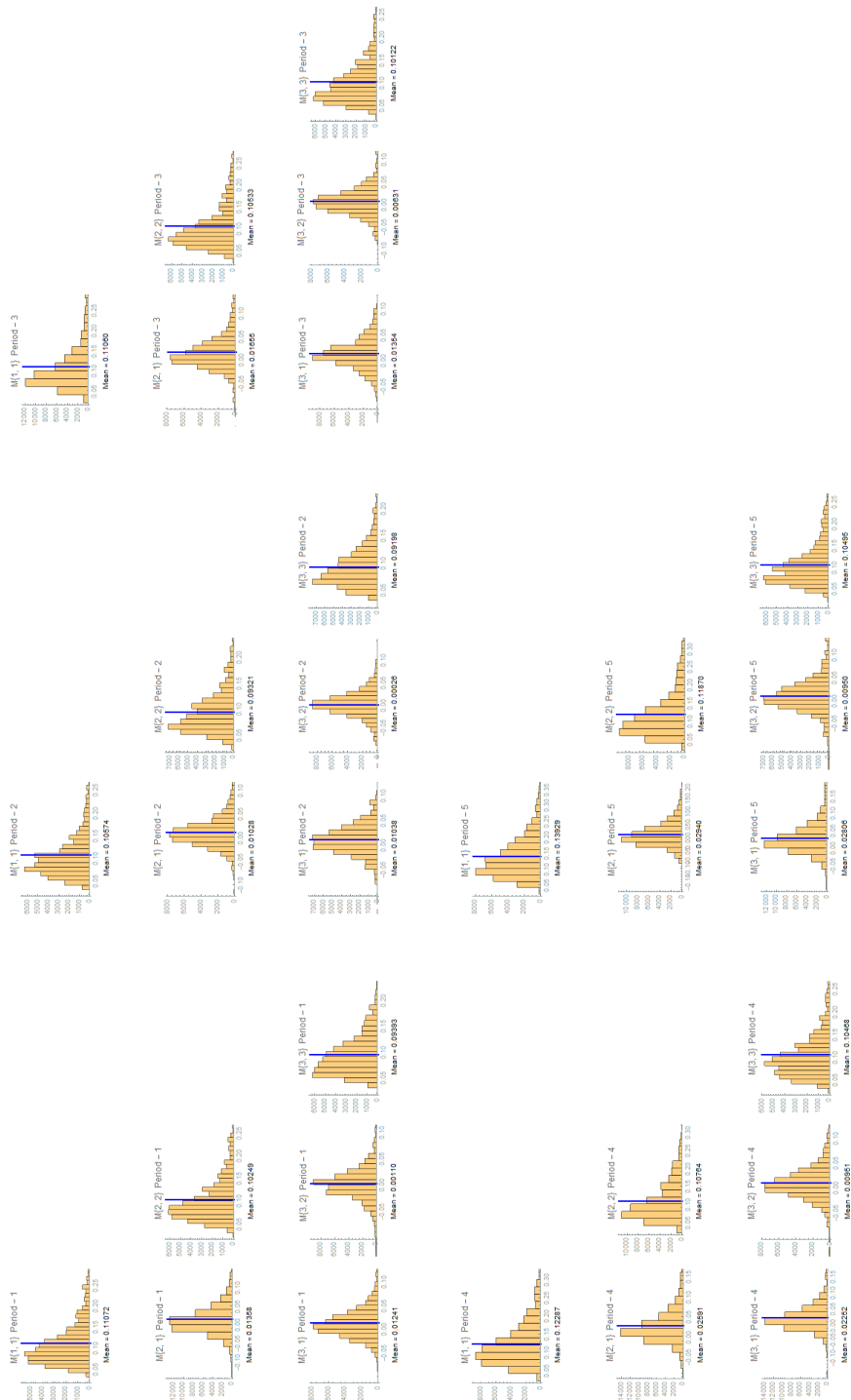


Figure 6.19: Histogram of M_1 componentwise

From Table 6.4, we are now able to state that component 2(Work Hours) has increased its correlated behaviour with component 1(Output), much more than in any of the previous simulation. However, there is indication of a trade-off, since at the same time, the correlation between Output and Consumption has decreased slightly from what we observed from the Matérn Smooth simulation.

Period 1	Period 2	Period 3
$\begin{pmatrix} 1 & & \\ 0.1275 & 1 & \\ 0.1217 & 0.0112 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & & \\ 0.1035 & 1 & \\ 0.1053 & 0.0029 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & & \\ 0.1533 & 1 & \\ 0.1280 & 0.0611 & 1 \end{pmatrix}$
Period 4	Period 5	
$\begin{pmatrix} 1 & & \\ 0.2253 & 1 & \\ 0.1986 & 0.0896 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & & \\ 0.2287 & 1 & \\ 0.2321 & 0.0851 & 1 \end{pmatrix}$	

Table 6.4: Correlation Matrices For Squared Exponential Covariance Function

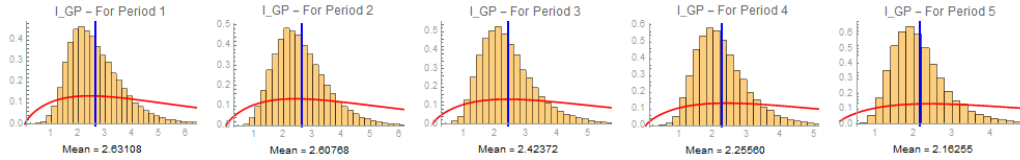


Figure 6.20: Histograms and Density Prior for l of the Covariance Function

We observe a similar behaviour to l_{gp} now, decreasing from 2.63 to 2.16 approximately. It tends now to rougher behaviours, i.e. smaller values of the length-scale, trying to compensate for the imposition of infinitely differentiable functions by the Squared Exponential Covariance function.

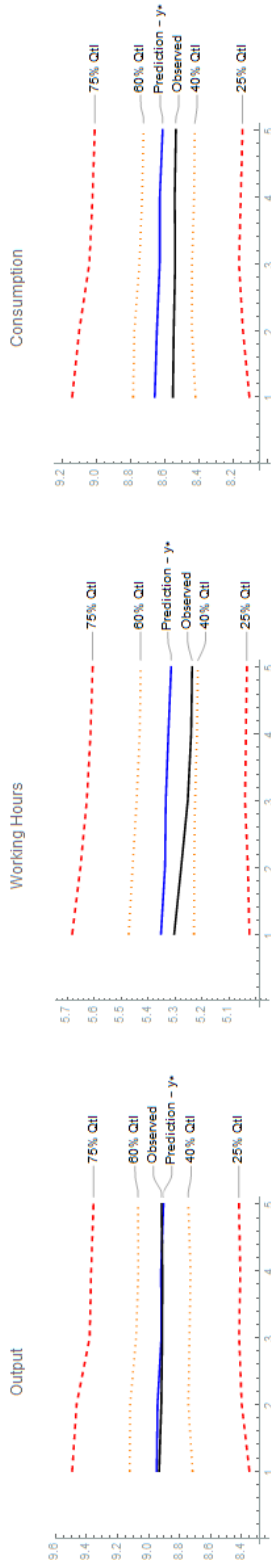


Figure 6.21: Posterior Predictive Intervals for y^*

136

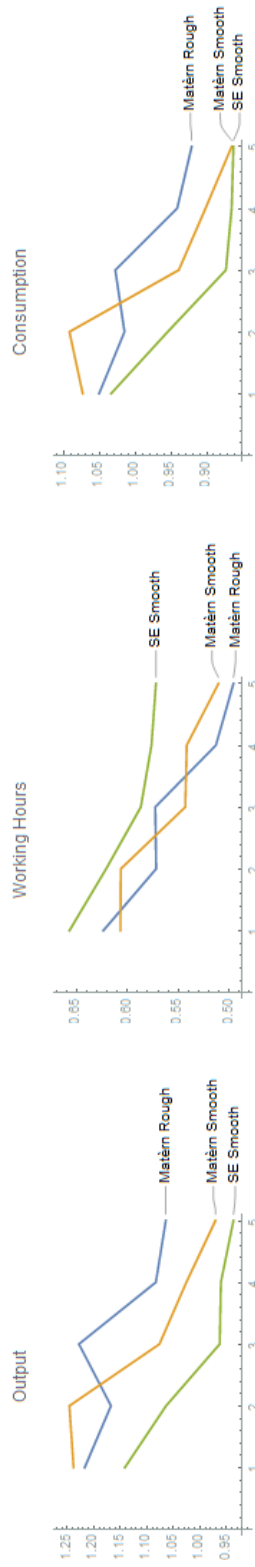


Figure 6.22: Posterior Predictive Intervals Length for y^*

In Figure 6.22 we see that the imposition of our new covariance function was enough to decrease the predictive interval magnitude, an observation which had already been hinted at with the histograms for the posteriors of the economic parameters. It seems that increasing the smoothness of our Covariance function has made our model to be surer of where future values of the economic quantities will be.

Another topic which could try to analyse using the results from all three simulations is the presence of non-identifiability issues. If there were no identifiability issue, as it was alluded in previous chapters, one would expect a marked tightening in the prediction intervals or posterior parameter distributions with the use of a smoother Covariance function and tighter priors. Further research on this issue is desirable, with a method that can deal with a greater amount of data and ensure more certain conclusions. However, the tightening of certain posteriors seems to give weight to the nonexistence of identifiability issues.

The resulting RMSE, (0.0149859, 0.0718813, 0.0923144) reflects the increased correlation of the Work Hours with Output, where we can observe a slight increase in predictive performance, but with a marked decrease for Consumption(component 3).

6.5 Conclusions

In this section, we applied our UQ methodology to the period of the great international crisis. One of our objectives was to compare the forecasting performance from the models exposed in Negro and Schorfheide (2013), with our GP discrepancy term. However, since the different models forecast quantities of interest of different nature, we decided for a very different criteria than the usual RMSE comparison from previous chapters. In this chapter, we decided to compare the magnitude of the predictive interval, and verify whether we could capture the crisis within certain bounds. Contrary to Negro and Schorfheide (2013), where the very complex economic models used had a very hard time to comprehend the economic variables during the crisis period, even for the 90% bounds, our methodology managed to capture them, for most periods even for the 20% interval prediction. Again, this is another example of a bitter sweet result. On one hand, it is very interesting to be able to capture the crisis, but on the other hand, the prediction intervals seem to be so broad as to encompass too many possibilities to be of practical use in some forecasting applications.

We used tighter priors, smoother covariance functions, but the issue still remained, albeit alleviated. More research is still needed to verify whether this is in fact a problem, and whether it can be solved by adding more data. Another topic we also touched was the presence of non-identifiability issues. When we have identifiability issues, very markedly different parameter values are fitted, which makes the posterior distributions very diffuse. By imposing tighter priors, we observed a tightening for the posterior predictive distributions, which could indicate that there is identifiability. A future research

venue should be directed toward a method that can deal with a greater amount of data, and may give a more certain answer to the (in)existence of non-identifiability issues.

Chapter 7

Conclusions for the Present and the Future

In the introductory chapters, we reviewed the state of the art of model misspecification correction, noticed their several different limitations, and indicated how our framework would contribute to it, by being applicable to any type of SSM, even non-linear, as a consequence of a black-box perspective, and how our discrepancy/bias term should be able to capture very complex dynamics. We also raised the possibility that our framework of accounting for model discrepancy could create some identifiability issues, and concerning that topic we postponed to chapter 6 a fuller analysis. A method which prevents this issue would be highly desirable, but so far the state of the art is suboptimal. The methods for correcting non-identifiability in GP regression are still in their infancy, with severe limitations on the allowed functional form of the mean function, and considerable computational burden. Therefore, future research on this area would be welcomed by the practitioner who is also interested in parameter estimation interpretability and not just in predictive accuracy.

In chapter 4, despite this issue, and as referred previously, in Chapter 3, the eventual presence of non-identifiability did not invalidate the method's predictive performance. Furthermore, the full GP showed a remarkable predictive accuracy, even if it was not enough to improve the learning of the model, which was similar that of Ireland (2004). The computational burden of our method only allowed us to use a small sample of observations for our simulations. This seemed to us the main bottleneck in the implementation of our method, and hence we decided the second chapter containing original research would be dedicated to improve this aspect.

Thus, the main objective of chapter 5 was to decrease the computational complexity of our UQ of model discrepancy framework by using an approximation to the GP implementing Hilbert Space techniques (HRRGP). The method did indeed decrease the computational time, although our implementation was found to be slightly lacking. As more observations are used, the difference in computational performance will tend to be increasingly favourable to the HRRGP, its absolute performance still seems to be prohibitively high. Using an approximation, we could already expect a trade-off between time and accuracy. Unfortunately, the time improvement was proportionally below than the worsening in forecasting performance. However, if more data were to be used, then this relation may very easily be inverted.

A surprising fact was that the HRRGP method had a deteriorated mixing for the economic parameters, when compared to the full GP method. A third venue of future research would be to modify our simple Random-Walk type of Metropolis-Hastings step. Even though it was not incorporated explicitly in our work, we did try a simple gradient approach to the MH step by using Matrix Calculus, but unfortunately severe numerical instability issues prevented an evaluation of some of the expressions in some iterations, calling into question the convergence validity of our algorithm. Therefore, we decided not to use this approach.

Another shortcoming of the HRRGP method, at least in an economics setting, is the great dependence on L_i (too big or too low would increase even more the RMSE). It would be interesting to investigate how severe this dependence is, and how should one choose the values. How does a discrimination for the values associated with the different components of x_t influence our results? This dependence on the L_i could also induce the researcher into 'overfitting' our GP, which is less than desirable.

An alternative to our HRRGP method, could be the use of an online algorithm. The use of an online algorithm as that in M. Cai et al. (2019), which itself already uses approximations to the likelihood, accumulated with the approximation of the HRRGP method, risks worsening the forecasting performance of our method or worse, defeat our purpose of quantifying model misspecification by adding even more sources of uncertainty.

In the 6th chapter, we delved into the performance of our UQ method for one of the most serious periods of our recent economic history, the great international financial crisis. Our main objective was in a sense to compare the forecasting performance from the models exposed in Negro and Schorfheide (2013), with that of a discrepancy term with a GP prior. However, due to the different nature of the quantities of interest to be forecasted, we constrained ourselves to the comparison of the magnitude of the predictive interval, and

verify whether we could capture the crisis within certain % bounds. Contrary to Negro and Schorfheide (2013), where the very complex economic models used had a very hard time comprising the variables during the crisis period even for the 90% bounds, our methodology managed to capture them, for most periods even for the 20% interval prediction. However, these prediction intervals may seem slightly too broad to be of practical use in forecasting. Even after using tighter priors, and smoother covariance functions, the issue still prevailed, albeit slightly alleviated. A venue of future research is to verify whether this problem is solved by adding more data. The possibility of being in the presence of non-identifiability issues gained a bit less traction with these exercises, since if we are in the presence of serious identification problems, one might expect tighter priors not to have any effect on the tightening of posteriors, which they did. Hence, at least for the RBC model tested in this work, non-identification seemed not to be that serious.

Regardless, this method looks promising, if not for the immediate present, then for the near future, where increasing computational capabilities seem boundless.

Appendix A

A Brief Introduction to Gaussian Processes

A.1 Univariate Gaussian Processes

There are two broad perspectives on how to see Gaussian Processes: the *weight-space* view, and the *function-space* view.

A.1.1 Function-Space View

One can view a GP from a function-space perspective.

Definition A.1.1.1 *A Stochastic Gaussian Process is a collection of r.v. such that any finite subcollection has a joint Gaussian distribution.*

Similarly to its sister distribution, we can completely determine a Gaussian Process for a $f(x)$ by defining its mean function $m(x) := E(f(x))$ and its covariance function $k(x, x') := E(f(x) - m(x))(f(x') - m(x'))$, and writing

$$f(x) \sim \mathcal{GP}(m(x), k(x, x'))$$

There are several possibilities for the choice of a covariance function $k(x, x')$ as we shall see later on.

Let us define the matrix $K(X, X') := [k(x^i, x'^j)]_{i,j}$. From the above, and $f_* = (f(x_*^1), \dots, f(x_*^N))$, we obtain

$$f_* \sim N \left(\begin{pmatrix} m(x_*^1) \\ \vdots \\ m(x_*^N) \end{pmatrix}, K(X_*, X_*) \right)$$

Let us consider the following regression model $y_i = f(x^i) + \epsilon_i$, where $\epsilon_i \sim^{iid} N(0, \sigma_n^2)$. Hence, we have

$$\text{cov}(y) = K(X, X) + \sigma_n^2 I$$

and thus,

$$\begin{pmatrix} y \\ f_* \end{pmatrix} \sim N \left(\begin{pmatrix} m_X \\ m_{X_*} \end{pmatrix}, \begin{pmatrix} K(X, X) + \sigma_n^2 I & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{pmatrix} \right)$$

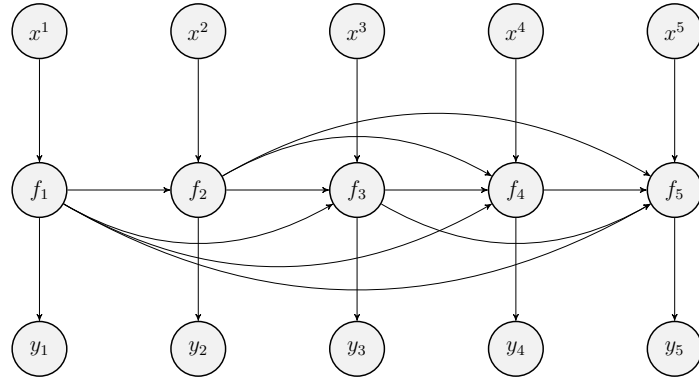


Figure A.1: Graphical Model for a GP Regression

In Figure A.1, we observe a graphical representation of Regression GP Model. One may notice that y_i s are independent when conditioned on f_i , but not when conditioned on x^i only.

Using the properties of the Normal distribution, we can derive the following predictive distribution

$$\begin{aligned} f_* | X, y, X_* &\sim N(m_*, K_*), \text{ where} \\ m_* &:= m_{X_*} + K(X_*, X)(K(X, X) + \sigma_n^2 I)^{-1}(y - m_X) \\ K_* &:= K(X_*, X_*) - K(X_*, X)(K(X, X) + \sigma_n^2 I)^{-1}K(X, X_*) \end{aligned}$$

A.1.2 Weight-Space View

In the weight-space view, we see a GP as way to project the inputs into a high dimensional space using the a set of *basis functions*, and then apply the linear model in this space instead of directly on the inputs. For now, assume the basis functions, independent of the weights w , are given.

Let the function $\phi(x)$ map a D_x -dimensional input into a D_ϕ -dimensional feature space. And let $\phi(X) = (\phi(x^1) \cdots \phi(x^{N'}))$ with $\{x^1, \dots, x^{N'}\}$ being the training set.

Consider a standard linear regression setting

$$y_i = f(x^i) + \epsilon_i$$

with $\epsilon_i \sim^{iid} N(0, \sigma_n^2)$

Now, instead of $f(x) = x^\top w$, assume $f(x) = \phi(x)^\top w$, where w has dimension D_ϕ , and prior $w \sim N(0, \Sigma_p)$. Using the Bayesian linear model regression formulas, and substituting $\phi(X)$ for X ¹, i.e. regressing on the basis functions instead of on the inputs, the predictive distribution now becomes

$$f_* \mid x_*, X, y \sim N \left(\frac{1}{\sigma_n^2} \phi(x_*)^\top A^{-1} \phi(X) y, \phi(x_*)^\top A^{-1} \phi(x_*) \right)$$

with $A = \sigma^{-2} \phi(X) \phi(X)^\top + \Sigma_p^{-1}$. Defining $K := \phi(X)^\top \Sigma_p \phi(X)$, we then have $\sigma_n^{-2} \phi(X) (K + \sigma_n^2 I) = \sigma_n^{-2} \phi(X) (\phi(X)^\top \Sigma_p \phi(X) + \sigma_n^2 I) = A \Sigma_p \phi(X)$.

It can be shown we can rewrite the above predictive equation into

$$f_* \mid x_*, X, y \sim N(\phi(x_*)^\top \Sigma_p \phi(X) (K + \sigma_n^2 I)^{-1} y, \phi(x_*)^\top \Sigma_p \phi(x_*) - \phi(x_*)^\top \Sigma_p \phi(x_*) (K + \sigma_n^2 I)^{-1} \phi(x_*)^\top \Sigma_p \phi(x_*))$$

We can notice that the feature space in the shape of $\phi(\cdot)^\top \Sigma_p \phi(\cdot)$, therefore we can rewrite again the above using $k(x, x') := \phi(x)^\top \Sigma_p \phi(x')$ ², which we shall call covariance function or kernel.

A.1.3 Covariance Functions

Even though the prior of a GP may have zero mean function, the posterior distribution will adapt to the data, and have a non zero mean. Therefore it is usually not very limitative to impose a prior GP with a zero mean. More important than the mean function for the GP is the covariance function.

The covariance functions will contain our most important assumptions on the function we wish to learn.

A *stationary* covariance function is a function of $x - x'$ only, being invariant to translations. If the covariance function is a function of $|x - x'|$, then it is called *isotropic*.

¹See Rasmussen and Williams formulas (2.9) and (2.11).

²This rewriting is called the kernel trick

Squared Exponential Covariance Function The *Squared Exponential*(SE) is of the form

$$k(x, x') = \exp\left(-\frac{(x - x')^2}{2l^2}\right)$$

where the parameter l is called characteristic length-scale. A GP with this covariance function will have mean-squared derivatives of all orders(infinitely mean-squared differentiable) and so, the functions it will learn are very smooth. This smoothness may be very restrictive in some applications. It is one of the most used covariance functions in the literature.

The Matèrn Class of Covariance Functions The Matèrn Class of covariance functions is given by

$$k(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{(2\nu)^{1/2}r}{l}\right)^\nu K_\nu\left(\frac{(2\nu)^{1/2}r}{l}\right)$$

where $r = |x - x'|$, and K_ν is a modified Bessel function. As $\nu \rightarrow \infty$ the GP process becomes infinitely smooth, and we obtain the squared exponential function. In fact, we have the following theorem

Theorem A.1.3.1 *With a Matèrn class, the process $f(x)$ is k -times mean-squared differentiable iff $\nu > k$.*

The Matèrn class of covariance functions has a simplified formula for $\nu = p + 1/2$ with $p \in \mathbb{N}_0$, and so the most used values are $\nu = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}$, where for $\nu = \frac{1}{2}$ the process is the roughest. For values greater than $\frac{7}{2}$, it becomes somewhat more difficult to distinguish between them.

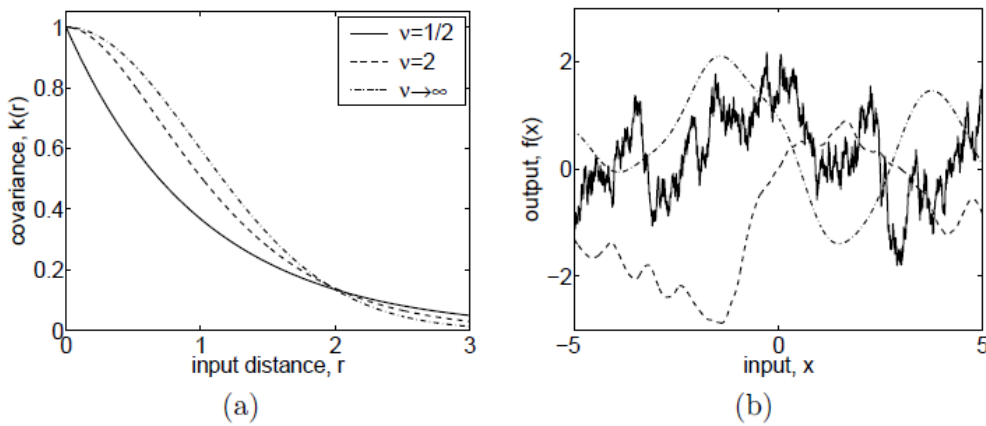


Figure A.2: Panel(a): Covariance Functions for different values of ν . Panel (b): Functions Randomly drawn from GP with respective Covariance Function in Panel (a). Source: Reference Rasmussen and Williams (2006)

A.1.4 Multi-Output Gaussian Process Regression

The following exposition follows closely H. Liu, J. Cai, and Ong (2018) and M. A. Alvarez, Rosasco, and Lawrence (2012). In a Multi-Output Gaussian Process Regression(MOGPR), we try to model a multidimensional function $f(x) \in \mathbb{R}^D$. Each of the D outputs $\{f_t\}_{t=1,\dots,D}$ may be considered simultaneously, or independently, depending on how the correlation function is defined. The objective of non-zero correlations is to use this inter-component dependence to improve over the inference of each of the D components considered separately.

Suppose we have the same number of observations (symmetry) for each component of f namely $X_d = x_{d,1}, \dots, x_{d,T}$, and exactly the same training set (isotopic) for each of them, which in our case, makes $X_d = x_1, \dots, x_T = \tilde{X} \in \mathbb{R}^{N \times T}$, where $x_{d,t} = x_t \in \mathbb{R}^N$.³

In the MOGPR, we assume that $f = (f_1, \dots, f_D)^\top$ follows a Gaussian Process of the form

$$f(x) \sim \mathcal{GP}(m(x), K(x, x'))$$

where the *multi-output* covariance function is a kernel for a *vector-valued* function, i.e., we can see $K(.,.)$ as being a *matrix-valued* function, which can be described by a *scalar* kernel receiving both inputs and input indices:

$$K_{d,d'}(x, x') = R((x, d), (x', d'))$$

where $K_{d,d'}(x, x') \in \mathbb{R}$ can be seen as the covariance between $f_d(x)$ and $f_{d'}(x')$, and

$$\text{Cov}(f(x), f(x')) = K(x, x') = \begin{bmatrix} K_{11}(x, x') & \dots & K_{1D}(x, x') \\ \vdots & \ddots & \vdots \\ K_{D1}(x, x') & \dots & K_{DD}(x, x') \end{bmatrix} \in \mathbb{R}^{D \times D}$$

In a regression of the form $y_d(x) = f_d(x) + \epsilon_d$, with $\epsilon_d \stackrel{iid}{\sim} N(0, \sigma_d^2)$ for $\sigma_d^2 \in \mathbb{R}$, we have to $p(y|f, x, \sigma) = N(y|f(x), \Sigma)$, with $\Sigma = \text{Diag}[\sigma_1^2, \dots, \sigma_D^2]$.

Also, given the following training set $X = (X_1, \dots, X_D) = (\tilde{X}, \dots, \tilde{X})$, it is helpful to define⁴

³We shall not use bold, because unless stated otherwise, we will always deal with multidimensional entities

⁴Here we diverge from the references cited above, since in the GP for State-Space models, the notation changes somewhat. In this literature we use $f_{1:T}$ just as defined above, meanwhile, in H. Liu, J. Cai, and Ong (2018) and M. A. Alvarez, Rosasco, and

$$K(X_*, X) = \begin{bmatrix} [K(x_{1*}, x_1)]_{D \times D} & \dots & [K(x_{1*}, x_T)]_{D \times D} \\ \vdots & \ddots & \vdots \\ [K(x_{T'_*}, x_1)]_{D \times D} & \dots & [K(x_{T'_*}, x_T)]_{D \times D} \end{bmatrix} \in \mathbb{R}^{T'_* D \times T D}$$

Thus, by definition of Gaussian Process, we have that

$$f_{1:T} := \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_T) \end{bmatrix} \sim N \left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_T) \end{bmatrix}, K(X, X) \right)$$

Let us also define

$$y_{1:T} := \begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix}$$

and some new data X_* , with $f_{1*:T'_*}$ being defined in the obvious way. It can be proved that

$$\begin{bmatrix} f_{1*:T'_*} \\ y_{1:T} \end{bmatrix} \sim N \left(\begin{bmatrix} m_{1*:T'_*} \\ m_{1:T} \end{bmatrix}, \begin{bmatrix} K(X_*, X_*) & K(X_*, X) \\ K(X, X_*) & K(X, X) + I_T \otimes \Sigma \end{bmatrix} \right)$$

since $\text{Cov}(f_{1*:T'_*}, y_{1:T}) = \text{Cov}(f_{1*:T'_*}, f_{1:T}) = K(X_*, X) \in \mathbb{R}^{DT'_* \times DT}$ and thus, from the formula for the conditional of a Multivariate Normal, we can show that

$$p(f_{*1:T} | X, y_{1:T}, X_*) = N(f_{*1:T} | \hat{f}_{*1:T}, \hat{K}_*)$$

5.

where

$$\hat{f}_{*1:T} = m_{*1:T} + K(X_*, X)(K(X, X) + I_T \otimes \Sigma)^{-1}(y_{1:T} - m_{1:T})$$

and

$$\hat{K}_* = K(X_*, X_*) - K(X_*, X)(K(X, X) + I_T \otimes \Sigma)^{-1}K(X, X_*)$$

Lawrence (2012), we would gather the components of f by observations instead, defining the vector $f_{1:D}$. For example if in one literature we define $K \otimes M$, in the other we will define it as $M \otimes K$.

⁵We are also conditioning on the parameters as Σ and those of the Gaussian Process Covariance function. They were not made explicit so as to not strain the notation even more

Appendix B

Bayesian Statistical Methods

B.1 Distributions for HRRGP

B.1.1 Matrix-Normal Distribution

In this section we will follow closely Gupta and Nagar (1999), with just some minor changes in notation.

The random matrix X with dimension $(n \times p)$ is said to follow the matrix normal distribution $\mathcal{MN}_{n,p}(M, U, V)$, if it can be represented by the following density:

$$p(X | M, U, V) = \frac{\exp\left(-\frac{1}{2} \text{Tr}\left[V^{-1}(X - M)^T U^{-1}(X - M)\right]\right)}{(2\pi)^{np/2} |V|^{n/2} |U|^{p/2}}$$

where Tr denotes trace and M is $(n \times p)$, U is $(n \times n)$ and V is $(p \times p)$.

The matrix normal can be related to the multivariate normal distribution in the following way:

Property B.1.1.1 $X \sim \mathcal{MN}_{n \times p}(M, U, V) \Leftrightarrow \text{vec}(X) \sim \mathcal{N}_{np}(\text{vec}(M), V \otimes U)$

Another way to relate both distributions, and help understand why U is the row covariance, and V is the column covariance are the following:

Property B.1.1.2 $X \sim \mathcal{MN}_{n \times p}(M, U, V) \Leftrightarrow X^\top \sim \mathcal{MN}_{p \times n}(M^\top, V, U)$

and

Property B.1.1.3 If $x_i \stackrel{\text{iid}}{\sim} \mathcal{N}_{p \times 1}(\mu, \Sigma)$, then

$$X = (x_1, \dots, x_n) \sim \mathcal{MN}_{p \times n}(M, U, V)$$

where $M = \mathbf{1}^\top \otimes \mu$, $U = \Sigma$ and $V = I_n$

Property B.1.1.4 If $X \sim \mathcal{MN}_{p \times n}(M, \Sigma, \Psi)$, with $D_{m \times p}$ and $\text{rank}(D) = m \leq p$, $C_{n \times t}$ and $\text{rank}(C) = t \leq n$, then

$$DXC \sim \mathcal{MN}_{m \times t}(DMC, D\Sigma D^\top, C^\top \Psi C)$$

From the above two last properties, we have an efficient way to sample from a matrix-normal using a simple multivariate normal.

Let us assume we want to draw a matrix \tilde{X} from $\mathcal{MN}(M, Q, V)$, then take $x_i \stackrel{\text{iid}}{\sim} \mathcal{N}_{p \times 1}(0, I_p)$. Then $X = (x_1, \dots, x_n) \sim \mathcal{MN}(0, I_p, I_n)$, and using the notation $\text{Chol}(M)^\top \text{Chol}(M) = M$, we thus obtain $\tilde{X} = M + \text{Chol}(Q)^\top X \text{Chol}(V) \sim \mathcal{MN}(M, \text{Chol}(Q)^\top \text{Chol}(Q), \text{Chol}(V)^\top \text{Chol}(V))$

B.1.2 Inverse-Wishart Distribution

A matrix X follows an Inverse-Wishart distribution, i.e. $X \sim \mathcal{IW}(\Psi, \nu)$, if its density function is:

$$f_X(X; \Psi, \nu) = \frac{|\Psi|^{\nu/2}}{2^{\nu p/2} \Gamma_p(\frac{\nu}{2})} |X|^{-(\nu+p+1)/2} \exp\left(-\frac{1}{2} \text{Tr}(\Psi X^{-1})\right)$$

where X and Ψ are $p \times p$ positive-definite matrices, and $\Gamma_p(\cdot)$ is the multivariate gamma function.

B.1.3 The Matrix-Normal Inverse-Wishart Conjugate Prior

The likelihood is

$$\log p(y_{1:T} \mid \Sigma, W) \propto -\frac{T}{2} \log(|\Sigma|) - \frac{1}{T} \text{Tr} \left(\sum_{t=1}^T (y_t - W\phi(x_t))^\top \Sigma^{-1} (y_t - W\phi(x_t)) \right)$$

Dealing with the expression inside the trace, and its properties, we have

$$\begin{aligned}
\text{Tr} \left(\sum_{t=1}^T (y_t - W\phi(x_t))^{\top} \Sigma^{-1} (y_t - W\phi(x_t)) \right) &= \sum_{t=1}^T \text{Tr} \left((y_t - W\phi(x_t))^{\top} \Sigma^{-1} (y_t - W\phi(x_t)) \right) \\
&= \sum_{t=1}^T \text{Tr} \left(\Sigma^{-1} (y_t - W\phi(x_t)) (y_t - W\phi(x_t))^{\top} \right) \\
&= \text{Tr} \left(\Sigma^{-1} \sum_{t=1}^T y_t y_t^{\top} - y_t (W\phi(x_t))^{\top} - (W\phi(x_t)) y_t^{\top} \right. \\
&\quad \left. + (W\phi(x_t)) (W\phi(x_t))^{\top} \right)
\end{aligned}$$

this last equality can be rewritten as

$$\text{Tr}(\Sigma^{-1}(\Upsilon - \Phi W^{\top} - W \Phi^{\top} + W \Psi W^{\top}))$$

where $\Upsilon = \sum_{t=1}^T y_t y_t^{\top}$, $\Phi = \sum_{t=1}^T y_t \phi(x_t)^{\top}$ and $\Psi = \sum_{t=1}^T \phi(x_t) \phi(x_t)^{\top}$.

For the priors, we have

$$\begin{aligned}
\log p(W, \Sigma) &= \log p(W \mid \Sigma) + \log p(\Sigma) \\
&\propto -\frac{1}{2}(D_y + l_{\Sigma} + 1) \log |\Sigma| - \frac{1}{2} \text{Tr}(\Sigma^{-1} \Lambda_{\Sigma}) \\
&\quad - \frac{1}{2} m \log |\Sigma| - \frac{1}{2} \text{Tr}(\Sigma^{-1} W V^{-1} W^{\top})
\end{aligned}$$

For the posterior we have,

$$\begin{aligned}
\log p(W, \Sigma \mid y_{1:T}, x_{1:T}) &= \log p(y_{1:T}, x_{1:T} \mid W, \Sigma) + \log p(W, \Sigma) \\
&\propto -\frac{1}{2}(D_y + m + l_{\Sigma} + T + 1) \log |\Sigma| - \frac{1}{2} \text{Tr}(\Sigma^{-1} \Lambda_{\Sigma}) \\
&\quad - \frac{1}{2} \text{Tr}(\Sigma^{-1}(\Upsilon - \Phi W^{\top} - W \Phi^{\top} + W(\Psi + V^{-1})W^{\top}))
\end{aligned}$$

Let us add $0 = -\frac{1}{2} \text{Tr}(\Sigma^{-1}(\Phi(\Psi + V^{-1})^{-1} \Phi^{\top} - \Phi(\Psi + V^{-1})^{-1} \Phi^{\top}))$ to above, and we get

$$\begin{aligned}
\log p(W, \Sigma \mid y_{1:T}, x_{1:T}) &\propto -\frac{1}{2}(D_y + m + l_\Sigma + T + 1) \log |\Sigma| \\
&\quad - \frac{1}{2} \text{Tr}(\Sigma^{-1}(\Lambda_\Sigma + \Upsilon - \Phi(\Psi - V^{-1})^{-1}\Phi^\top)) \\
&\quad - \frac{1}{2} \text{Tr}(-\Phi W^\top \Sigma^{-1} - \Sigma^{-1} W \Phi^\top + W^\top \Sigma^{-1} W (\Psi + V^{-1}) \\
&\quad + \Sigma^{-1} \Phi (\Psi + V^{-1})^{-1} \Phi^\top)
\end{aligned}$$

The 2nd trace, using the symmetry of some matrices, can be rewritten as

$$\begin{aligned}
& - \frac{1}{2} \text{Tr}(-W^\top \Sigma^{-1} \Phi - \Phi^\top \Sigma^{-1} W (\Psi + V^{-1})^{-1} + W^\top \Sigma^{-1} W \\
& + \Sigma^{-1} \Phi (\Psi + V^{-1})^{-1} \Phi^\top (\Psi + V^{-1})^{-1} (\Psi + V^{-1})) \\
& = -\frac{1}{2} \text{Tr}[(W - \Phi(\Psi + V^{-1})^{-1})^\top \Sigma^{-1} (W - \Phi(\Psi + V^{-1})^{-1}) (\Psi + V^{-1})]
\end{aligned}$$

which results in the posterior distributions

$$W \mid \Sigma \sim \mathcal{MN}(\Phi(\Psi + V^{-1})^{-1}, \Sigma, (\Psi + V^{-1})^{-1})$$

and

$$\Sigma \sim \mathcal{IW}(l_\Sigma + T, \Lambda_\Sigma + \Upsilon - \Phi(\Psi + V^{-1})^{-1} \Phi^\top)$$

B.2 Introducing Particle Filtering Algorithms

Particle Filtering refers to a class of algorithms where parallel chains (the so called particles) are splitted, killed or created with the objective of exploring with greater detail the space where the target distribution has more mass. These type of algorithms may also be viewed as a specific instance of a Sequential Monte Carlo (SMC) algorithm. The exposition here is based on several references such as Andrieu, Doucet, and Holenstein (2010), Cappé, Godsill, and Moulines (2007), Doucet and Freitas (2001), and Doucet and Johansen (2008)

B.2.1 Monte Carlo Method

The Monte Carlo method approximates a general probability density $p_n(x_{1:n})$, by the empirical measure¹

¹An empirical measure μ is of the form $\mu = \frac{1}{N} \sum_{i=1}^N \delta_{X_i}$

$$\hat{p}_n(x_{1:n}) = \frac{1}{N} \sum_{i=1}^N \delta_{X_{1:n}^i}(x_{1:n})$$

where $X_{1:n}^i \sim p_n(x_{1:n})$.

In this way,

$$\int \phi(x_{1:n}) p_n(x_{1:n}) dx_{1:n} \approx \int \phi(x_{1:n}) \hat{p}_n(x_{1:n}) dx_{1:n} = \frac{1}{N} \sum_{i=1}^N \phi(X_{1:n}^i)$$

Intuitively, what is written above is draw $X_{1:n}$ from p_n , and then approximate $E(\phi(X_{1:n}))$ by using $\frac{1}{N} \sum_{i=1}^N \phi(X_{1:n}^i)$.

So, far we have been assuming that we can sample from $p_n(x_{1:n})$. However, specially for greater dimensional spaces, this becomes impossible, and so we might want to use what is called Importance Sampling.

B.2.2 Importance Sampling

The Importance Sampling (IS) method requires the use of an importance density function $q_n(x_{1:n})$ such that $p_n(x_{1:n}) > 0 \Rightarrow q_n(x_{1:n}) > 0$, and from which it is relatively easy to sample from. We will assume to know the functional form of p_n, q_n , but not how to draw efficiently, or at all, from p_n .

We have

$$p_n(x_{1:n}) = \frac{w_n(x_{1:n}) q_n(x_{1:n})}{Z_n}$$

with $Z_n = \int w_n(x_{1:n}) q_n(x_{1:n}) dx_{1:n}$, and $w_n(x_{1:n}) = \frac{p_n(x_{1:n})}{q_n(x_{1:n})}$ being called the unnormalized weight function. Depending on our context, we may be interested in a density such as $p_n(x_{1:n}, y_{1:n})$ instead.

where the X_n are random variables $(\Omega, \mathcal{F}) \rightarrow (X, \mathcal{F}_X)$. For any $\omega \in \Omega$, and $B \in \mathcal{F}_X$, we have

$$\mu(\omega)(B) = \frac{1}{N} \sum_{i=1}^N \delta_{X_i(\omega)}(B)$$

where

$$\delta_{X_i(\omega)}(B) = \begin{cases} 1 & : X_i(\omega) \in B \\ 0 & : X_i(\omega) \notin B \end{cases}$$

and for fixed ω , we can interpret $\mu(\omega)$ as a measure. Furthermore, if $f : X \rightarrow \mathbb{R}$, then

$$\mu f(\omega) := \int_X f d\mu(\omega) = \frac{1}{N} \sum_{i=1}^N \int_X f d\delta_{X_i(\omega)} = \frac{1}{N} \sum_{i=1}^N f(X_i(\omega))$$

Now, sampling $X_{1:n}^i \sim q_n(x_{1:n})$ and if we use an empirical approximation to $q_n(x_{1:n})$, i.e. $\hat{q}_n(x_{1:n}) = \frac{1}{N} \sum_{i=1}^N \delta_{X_{1:n}^i}(x_{1:n})$, we will obtain

$$\hat{Z}_n = \frac{1}{N} \sum_{i=1}^N w_n(X_{1:n}^i)$$

and therefore

$$\hat{p}_n(x_{1:n}) = \sum_{i=1}^N W_n(x_{1:n}) \delta_{X_{1:n}^i}(x_{1:n})$$

with $W_n(x_{1:n}) = \frac{w_n(x_{1:n})}{\hat{Z}_n}$, since we must have $\int \hat{p}_n(x_{1:n}) dx_{1:n} = 1$.

Intuitively, what is stated above is to draw $X_{1:n}$ from $q_n(x_{1:n})$ and approximate $E(\phi(X_{1:n}))$ using $\frac{1}{N} \sum_{i=1}^N \phi(X_{1:n}^i) W_n(X_{1:n}^i)$.

A good proposal would be one which resembles $p_n(x_{1:n})$ so that it minimizes the variance of the importance weights w_n .

One of the problems with IS is that it is not adequate for recursive estimation. In the context for example of a State-Space model, we may be interested in estimating $p_n(x_{1:n}) = p(x_{1:n} | y_{1:n})$, and the marginal posterior $p(x_n | y_{1:n})$, the latter also known as the *filtering distribution*². Then, when we have a new y_{n+1} available, we must recompute the $w(x_{1:n+1})$ from the whole $x_{1:n+1}$ sequence, which may easily become computationally too heavy.

B.2.3 Sequential Importance Sampling

The Sequential Importance Sampling technique was created specifically to correct the limitation of IS stated above. Henceforth, we will now restrict ourselves to the specific context of State-Space Models³(SSM) which is the one we are interested in, although Sequential Monte Carlo(SMC) algorithms could be viewed from a more general perspective, which the interested reader may consult in reference Doucet and Johansen (2008).

Let us consider the following SSM:

$$\begin{aligned} x_1 &\sim p(x_1) \\ x_t | x_{t-1} &\sim p_x(x_t | x_{t-1}) \\ y_t | x_t &\sim p_y(y_t | x_t) \end{aligned}$$

²The problem of filtering may be described as that of finding $p(x_n | y_{1:n})$, whereas that of smoothing is to find $p(x_t | y_{1:n})$ for some $t < n$, and prediction is that of finding $p(x_t | y_{1:n})$ for some $t > n$.

³For some authors, the difference between what is a State-Space Model, a Bayesian Dynamic Model, or a Hidden Markov Model is a 'null set'(negligible)

Our objective is to find an approximation to $p(x_{1:n} \mid y_{1:n})$. It is usual to select a proposal distribution of the form

$$q_n(x_{1:n} \mid y_{1:n}) = q_1(x_1) \prod_{i=2}^n q_i(x_i \mid x_{1:i-1}, y_{1:i})$$

which factorises in similar way to our target distribution. To compute sequentially the unnormalized weights for the empirical approximation, we should notice the following:

$$\begin{aligned} p(x_{1:n} \mid y_{1:n}) &= \frac{p(y_n \mid x_{1:n}, y_{1:n-1})p(x_{1:n} \mid y_{1:n-1})}{p(y_n \mid y_{1:n-1})} \\ &= \frac{p(y_n \mid x_n)p(x_n \mid x_{1:n-1}, y_{1:n-1})p(x_{1:n-1} \mid y_{1:n-1})}{p(y_n \mid y_{1:n-1})} \\ &= p(x_{1:n-1} \mid y_{1:n-1}) \frac{p(y_n \mid x_n)p(x_n \mid x_{n-1})}{p(y_n \mid y_{1:n-1})} \end{aligned}$$

This is what is called the *recursive Bayesian estimation* formula. Henceforth, for easiness of understanding, we will simplify the notation, by not distinguishing between X, Y and x, y . In the definition of $w(x_{1:n}^i)$, let us incorporate our target distribution $p_n(x_{1:n}) = p(x_{1:n} \mid y_{1:n})$ ⁴, so that we obtain the following

$$\begin{aligned} w(x_{1:n}^i) &= \frac{p(x_{1:n-1}^i \mid y_{1:n-1})}{q_{n-1}(x_{1:n-1}^i \mid y_{1:n-1})} \frac{p(y_n \mid x_n^i)p(x_n \mid x_{n-1}^i)}{q_n(x_n^i \mid x_{1:n-1}^i, y_{1:n})p(y_n \mid y_{1:n-1})} \\ &= w(x_{1:n-1}^i) \frac{p(y_n \mid x_n^i)p(x_n \mid x_{n-1}^i)}{q_n(x_n^i \mid x_{1:n-1}^i, y_{1:n})p(y_n \mid y_{1:n-1})} \\ &\propto w(x_{1:n-1}^i) \frac{p(y_n \mid x_n^i)p(x_n \mid x_{n-1}^i)}{q_n(x_n^i \mid x_{1:n-1}^i, y_{1:n})} \end{aligned}$$

At each iteration of this step, we can obtain empirical estimates of p_n and Z_n , i.e. \hat{p}_n and \hat{Z}_n .

A common choice of proposal is to use $q_n(x_{1:n} \mid y_{1:n}) = p_x(x_{1:n})$, resulting in what we will see below as *particle filters*. A particle is each i -th draw x_n^i , and to $x_{1:n}^i$ we call the particle's trajectory.

Note that we can interpret the SIS as an instance of SI, and it is known that for high-dimensional spaces, i.e. in our context a high-value of n , this method is inefficient due to the great number of particles it will require.

⁴In Doucet and Johansen (2008), the reader may see how to incorporate other more general target distributions

The problem with SIS framework, a reflection of the issue just described, is that as n increases, the distribution \hat{p}_n becomes more and more easily degenerate, i.e. less and less $w(x_{1:n}^i)$ will have non-zero value after some recursions. This degeneration prevents us from finding a good approximation of p_n . To alleviate this issue, usually one uses resampling.

Resampling

The idea of resampling is to eliminate the trajectories with smaller weights, and multiply those with greater weights.

With that purpose in mind, one of the most intuitive ways to resample is to simply choose $X_{1:n}^i$ with probability W_n^i , i.e. we are sampling from a previous sample (resampling) drawn from \hat{p}_n , favouring those particles with greater normalised weights. Hence, a natural way is to do a Multinomial resampling, even though the literature has shown the existence of more efficient resampling methods.

One of the most important advantages of resampling is to allow us to eliminate particles $X_{1:n}^i$ ⁵ with a low weight/probability and favour those with higher probability. The intuition on the desirability for this is simple, we do not want to have just few particles to represent regions of high probability mass, instead we desire to explore these regions well, in order to find a good approximation to p_n .

Furthermore, resampling helps to understand the notion of *path* degeneracy, i.e. after a number of iterations the number of different segments $x_{1:k}$ for $k \ll n$ is reduced. If we resample from previous particles, we are reducing the number of different values representing $x_{1:k}$. However, this problem of *path* degeneracy is a consequence of *weight* degeneracy, originating in SI methods, since we are trying to estimate a distribution on a space of increasingly arbitrarily high dimension in n . After resampling, the $W_n(X_{1:n}^i)$ are reset to $1/N$.

B.2.4 Particle Filtering

Particle filtering algorithms can be seen as a combination of Sequential Importance Sampling plus a Resampling step for the purpose of finding an approximation to $p(x_{1:n} | y_{1:n})$ or $p(x_n | y_{1:n})$.

One of the simplest particle filters is the Bootstrap Filter, which uses $q_n(x_{1:n} | y_{1:n}) = p_x(x_{1:n})$. In the following, we use $w_n^i := w_n(X_{1:n}^i)$, and in each step where we the index i shows, we do for $i = 1, \dots, N$.

⁵In some literature, for a frugal nomenclature, authors use particle to refer to a particle or a particle's trajectory.

Algorithm 8 Bootstrap Filter

Input: $y_{1:T}$, N (sample size)

Output: $x_{1:T}^i \sim p(x_{1:T} \mid y_{1:T})$

- 1: Draw $x_1^i \sim p_x(x_1)$
 - 2: Set $W_n^i = \frac{1}{N}$
 - 3: **for** $t \in 2, \dots, T$ **do**
 - Importance Sampling Step
 - 4: Draw $x_t^i \sim p_x(x_t \mid x_{t-1}^i)$
 - 5: Set $x_{1:t}^i := (x_{t-1}^i, x_t^i)$
 - 6: Compute $w_t^i = p_y(y_t \mid x_t^i)$
 - 7: Normalise w_t^i
 - Resampling Step
 - 8: Draw $\tilde{x}_{1:t}^i \sim \text{Mult}(x_{1:t}^i, W_t^i : i = 1, \dots, N)$
 - 9: Set $x_{1:t}^i = \tilde{x}_{1:t}^i$
 - 10: Set $w_t^i = \frac{1}{N}$
 - 11: **end for**
 - 12: Draw k with $P(k = i) \propto w_T^i$
 - 13: **return** $x_{1:T}^k = x_{1:T}^k$
-

Although for this filter, Step 10 is superfluous, as is Step 2, since in Step 6 we are already imposing $w_t^i = \frac{1}{N}$, for more general particle filters these steps may not be.

Another usual notation in the literature is that of A_{t-1}^k being the index of the 'ancestor' particle at time $t - 1$ of particle $X_{1:t}^k$, and B_t^k the index so that $X_{1:t}^k = (X_1^{B_1^k}, X_2^{B_2^k}, \dots, X_t^{B_t^k})$.

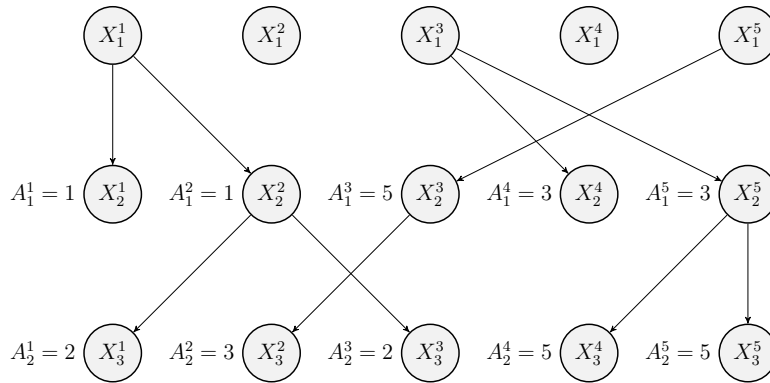


Figure B.1: An example of particles generated by an SMC algorithm

From Figure B.1, we can see that the first particle is $X_{1:3}^1 = (X_1^1, X_2^2, X_3^1)$ with $B_{1:3}^1 = (1, 2, 1)$.

Particle Gibbs Sampler

For Particle Filtering, when using a Gibbs sampler methodology, it is necessary to use a conditional SMC update. This SMC, contrary to the previous SMCs, takes as given a predefined particle path $X_{1:T}$ with respective $B_{1:T}$ ancestral lineage, which is determined to survive, regardless of the remaining particles created as in the usual SMC. This algorithm was first presented in Andrieu, Doucet, and Holenstein (2010), to the best of our knowledge.

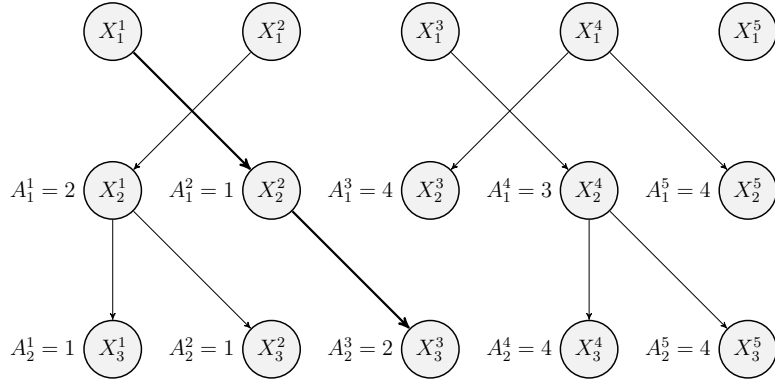


Figure B.2: An example of particles generated by a Conditional SMC algorithm

In the Figure B.2, the reference particle was fixed at $X_{1:3}^3 = (X_1^1, X_2^2, X_3^3)$, in bold arrows.

Particle Gibbs with Ancestor Sampling: PGAS

This relatively new method tries to tackle the problem of path degeneracy by doing a small modification to the usual Particle Gibbs (PG). This algorithm was first introduced in Lindsten, Jordan, and Schön (2014).

Algorithm 9 PGAS Markov Kernel

Input: Reference Trajectory $x'_{1:T}$

Output: $x_{1:T} \sim p_\theta(x_{1:T} \mid y_{1:T})$

- 1: Draw $x_1^i \sim r_{\theta,1}(x_1)$ for $i = 1, \dots, N - 1$
 - 2: Set $x_1^N = x'_1$
 - 3: Set $w_1^i = W_{\theta,1}(x_1^i)$ for $i = 1, \dots, N$
 - 4: **for** $t \in 2, \dots, T$ **do**
 - 5: Draw $\{A_t^i, x_t^i\} \sim M_{\theta,t}(A_t, x_t)$ for $i = 1, \dots, N - 1$
 - 6: Set $x_t^N = x'_t$
 - 7: Compute $\{\tilde{w}_{t-1|T}^i\}_{i=1}^N$ (see formulas below)
 - 8: Draw A_t^N with $P(A_t^N = i) \propto \tilde{w}_{t-1|T}^i$
 - 9: Set $x_{1:t}^i = (x_{1:t-1}^{A_t^i}, x_t^i)$ for $i = 1, \dots, N$
 - 10: Set $w_t^i = W_{\theta,t}(x_{1:t}^i)$ for $i = 1, \dots, N$
 - 11: **end for**
 - 12: Draw k with $P(k = i) \propto w_T^i$
 - 13: **return** $x_{1:T} = x_{1:T}^k$
-

where

$$M_{\theta,t}(A_t, x_t) = \frac{w_{t-1}^{A_t}}{\sum_l w_{t-1}^l} r_{\theta,t}(x_t \mid x_{1:t-1}^{A_t})$$

$$W_{\theta,t}(x_{1:t}) := \frac{p_{\theta,t}(x_{1:t})}{p_{\theta,t-1}(x_{1:t-1}) r_{\theta,t}(x_t \mid x_{1:t-1})}$$

$$\tilde{w}_{t-1|T}^i := w_{t-1}^i \frac{p_{\theta,t}((x_{1:t-1}^i, x'_{t:T}))}{p_{\theta,t-1}(x_{1:t-1}^i)}$$

In the formulas above $r_{\theta,n}$ is the proposal density that was denoted previously by q_n ⁶, and the $p_{\theta,t}(x_{1:t})$ function in our context is $p_\theta(x_{1:t}, y_{1:t})$.

A few words are in order to connect the previous sections in this appendix to PGAS. For the most observant reader, the formula for our un-normalised weight function may seem odd, i.e. for $W_{\theta,t}(x_{1:t})$.

Let us first observe that

$$p_\theta(x_{1:t}, y_{1:t}) \propto p_\theta(x_{1:t} \mid y_{1:t})$$

⁶Here I use r to make it clearer its connection with the main chapters of this thesis, by avoiding some confusion from the proposal densities used in the Metropolis-Hasting step in the algorithm specified in the main text

and so we have

$$W_{\theta,t}(x_{1:t}) \propto \frac{p_{\theta}(x_{1:t} \mid y_{1:t})}{p_{\theta}(x_{1:t-1} \mid y_{1:t})q_t(x_t \mid x_{1:t-1})}$$

To be coherent with the previous section on SIS, we would expect the denominator in the weight function to be the proposal from which we draw the $x_{1:t}^i$, and in fact, with the occurrence of resampling it is the case. To understand why, assume we already have some desired weights. With these weights, when we do the resampling, the $x_{1:t-1}^i$ will be drawn now from the empirical approximation $\hat{p}_{n-1}(x_{n-1}) = \hat{p}(x_{n-1} \mid y_{n-1})$ and then the particle is propagated using $q_t(x_t \mid x_{1:t-1})$, resulting in a proposal of the form $q_t(x_{1:t}) = \hat{p}_{t-1}(x_{1:t-1})q_t(x_t \mid x_{1:t-1})$, instead of the usual $q_t(x_{1:t}) = q_{t-1}(x_{1:t-1})q_t(x_t \mid x_{1:t-1})$. Now, we can use $q_t(x_{1:t}) = p_{t-1}(x_{1:t-1})q_t(x_t \mid x_{1:t-1})$ as the approximative proposal, and it is proved in references Andrieu, Doucet, and Holenstein (2010) and Lindsten, Jordan, and Schön (2014) that indeed the algorithm converges to the desired target distribution. In the above algorithm resampling and propagation are represented in the same step, with the use of the markov kernel $M_{\theta,t}$.

When compared to the simple Particle Gibbs, the PGAS has one simple extra step, namely step 8, where we sample the ancestor index of the reference particle, whereas the PG algorithm would just define $A_t^N = N$. This small addition to the algorithm has considerable positive effect on the quality of the mixing, allowing for a smaller number of particles to be used without incurring into such a severe path degeneracy, as can be observed in Lindsten, Jordan, and Schön (2014).

The formula for $\tilde{w}_{t-1|T}^i$ then can be seen as an instance of the Bayes theorem where w_{t-1}^i is the prior probability for the particle $x_{1:t-1}^i$, and the ratio of target densities $p_{\theta,t}$ as the likelihood of $x'_{t:T}$ given particle $x_{1:t-1}^i$.

B.3 Using a PGAS for y_M

The model with $b(x_t) = 0$ is

$$\begin{aligned} x_t &= Ax_{t-1} + \eta_t \\ y_t &= C_* + Cx_t + \epsilon_t \end{aligned}$$

The likelihood of such model is

$$\begin{aligned}
p(x_{1:T}, y_{1:T} \mid \theta) &= p(x_1) \prod_{t=2}^T p(x_t \mid x_{t-1}, A, Q) \cdot \prod_{t=1}^T p(y_t \mid x_t, C_*, C, \Sigma) \\
&= p(x_1) N(x_{2:T} \mid (I_{T-1} \otimes A)x_{1:T-1}, I_{T-1} \otimes Q) \\
&\quad \cdot N(y_{1:T} \mid C_{*1:T} + (I_T \otimes C)x_{1:T}, I_T \otimes \Sigma)
\end{aligned}$$

The weights for the state-space estimation algorithm become, in this context,

$$\begin{aligned}
\tilde{w}_{t-1|T}^i &= w_{t-1}^i \frac{p_\theta((x_{1:t-1}^i, \tilde{x}_{t:T}), y_{1:T})}{p_\theta(x_{1:t-1}^i, y_{1:t-1})} \\
&= w_{t-1}^i p(\tilde{x}_t \mid x_{t-1}^i, A, Q) \prod_{j=t+1}^T p(\tilde{x}_j \mid \tilde{x}_{j-1}, A, Q) \cdot \prod_{j=t}^T p(y_j \mid \tilde{x}_j, C_*, C, \Sigma) \\
&\propto w_{t-1}^i p(\tilde{x}_t \mid x_{t-1}^i, A, Q)
\end{aligned}$$

and

$$\begin{aligned}
w_t^i &= \frac{p_\theta(x_{1:t}^i, y_{1:T})}{p_\theta(x_{1:t-1}^i, y_{1:t-1}) p_\theta(x_t^i \mid x_{1:t-1}^i)} \\
&= p(y_t \mid x_t^i, C, C_*, \Sigma)
\end{aligned}$$

For predicting y_* we could have used similar formulas as in section 2.2 such as $p(y_* \mid x_*, y_{1:T}) \approx \frac{1}{N} \sum_{i=1}^N N(y_* \mid C_*[i], C[i], x_*)$.

Instead we decided for the more consensual $y_* = \bar{C}_* + \bar{C}x_*$.

B.3.1 Trace Plots

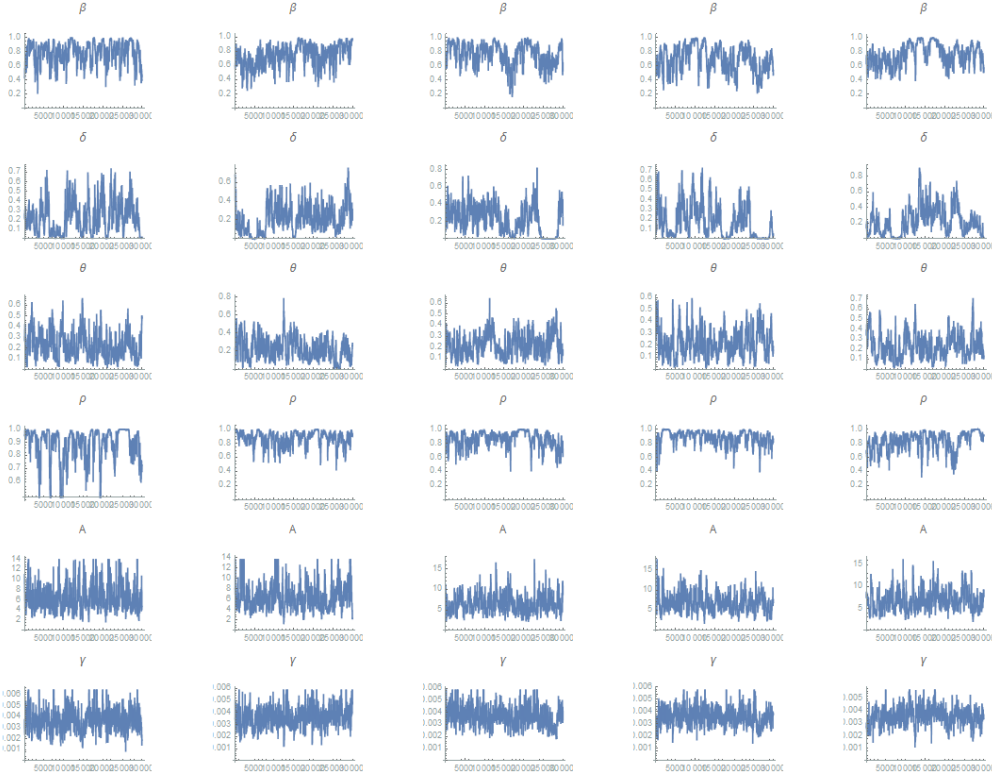


Figure B.3: Econ Trace for y_M

Figure B.3, was obtained using the same initial values, and remaining input arguments (s_1, s_2 , etc) as in the simulations above, in the main text. From the figure we can see that there is a greater difficulty for the chain to attain a proper mixing. However, it does not seem impede us from using the estimations obtained from running the algorithm.

B.4 The PGAS Markov Kernel for a HRRGP

Algorithm 10 PGAS Markov Kernel

Input: $x_{1:T}[n-1]$, $\theta[n-1]$, and N_p which is the number of particles

Output: $x_{1:T}[n]$

- 1: Set $\tilde{x}_{1:T} = x_{1:T}[n-1]$ as the reference trajectory
- 2: Draw $x_1^i \sim p(x_1 | \theta[n-1])$ for $i = 1, \dots, N_p - 1$.
- 3: Set $x_1^{N_p} = \tilde{x}_1$
- 4: Set $w_1^i = \frac{1}{N_p}$ for $i = 1, \dots, N_p - 1$.
- 5: **for** $t = 2, \dots, T$ **do**
- 6: Draw a_t^i with $P(a_t^i = j) \propto w_{t-1}^j$ for $i = 1, \dots, N_p - 1$.
- 7: Draw $x_t^i \sim p(x_t | x_{1:t-1}^{a_t^i}, \theta[n-1])$ for $i = 1, \dots, N_p - 1$.
- 8: Compute $\{\tilde{w}_{t-1|T}^i\}_{i=1}^{N_p}$ using

$$\tilde{w}_{t-1|T}^i = w_{t-1}^i \frac{p_{\theta[n-1]}((x_{1:t-1}^i, \tilde{x}_{t:T}), y_{1:T})}{p_{\theta[n-1]}(x_{1:t-1}^i, y_{1:t-1})}$$

- 9: Draw $a_t^{N_p}$ with $P(a_t^{N_p} = j) \propto \tilde{w}_{t-1|T}^j$
 - 10: Set $x_t^{N_p} = \tilde{x}_t$
 - 11: Set $x_{1:t}^i = (x_{1:t-1}^{a_t^i}, x_t^i)$ for $i = 1, \dots, N_p$.
 - 12: Compute $w_t^i = \frac{p_{\theta[n-1]}(x_{1:t}^i, y_{1:t})}{p_{\theta[n-1]}(x_{1:t-1}^i, y_{1:t-1}) \cdot r_{\theta[n-1]}(x_t^i | x_{1:t-1}^i)}$ for $i = 1, \dots, N_p$.
 - 13: **end for**
 - 14: Sample k with $P(k = i) \propto w_T^i$ and set $x_{1:T}[n] = x_{1:T}^k$
-

In Step 7, the density simplifies considerably to

$$p(x_t | x_{1:t-1}^{a_t^i}, \theta[l-1]) = N(A[l-1]x_{t-1}^{a_t^i}, Q[l-1]).$$

In step 8, the formula for the sequential computing of the weight for the particles can be simplified considerably by noticing

$$\begin{aligned} \tilde{w}_{t-1|T}^i &= w_{t-1}^i \frac{p_{\theta[n-1]}((x_{1:t-1}^i, \tilde{x}_{t:T}), y_{1:T})}{p_{\theta[n-1]}(x_{1:t-1}^i, y_{1:t-1})} \\ &= w_{t-1}^i p(\tilde{x}_t | x_{t-1}^i, \theta_x) \prod_{j=t}^T p(y_j | \tilde{x}_j, \theta_y) \\ &\propto w_{t-1}^i N(\tilde{x}_t | A[n-1]x_{t-1}^i, Q[n-1]) \end{aligned}$$

where the proportional symbol is used to indicate a missing factor which does not depend on the index i , i.e. it's a constant along all the particles for a given iteration in Algorithm 4.

In Step 12, we have $r_\theta(x_t|x_{t-1})$ which is a chosen proposal density. If for this proposal we use the transition density, then we have

$$\frac{p_{\theta[n-1]}(x_{1:t}^i, y_{1:t})}{p_{\theta[n-1]}(x_{1:t-i}^i, y_{1:t-1}) \cdot r_{\theta[n-1]}(x_t^i|x_{1:t-i}^i)} = N(y_t | W[n-1]\phi(x_t^i), \Sigma[n-1])$$

This will result in the following simplified algorithm:

Algorithm 11 PGAS Markov Kernel

Input: $x_{1:T}[n-1]$, $\theta[n-1]$, and N_p which is the number of particles

Output: $x_{1:T}[n]$

- 1: Set $\tilde{x}_{1:T} = x_{1:T}[n-1]$ as the reference trajectory
- 2: Draw $x_1^i \sim p(x_1|\theta[n-1])$ for $i = 1, \dots, N_p - 1$.
- 3: Set $x_1^{N_p} = \tilde{x}_1$
- 4: Set $w_1^i = \frac{1}{N_p}$ for $i = 1, \dots, N_p - 1$.
- 5: **for** $t = 2, \dots, T$ **do**
- 6: Draw a_t^i with $P(a_t^i = j) \propto w_{t-1}^j$ for $i = 1, \dots, N_p - 1$.
- 7: Draw $x_t^i \sim N(A[l-1]x_{t-1}^{a_t^i}, Q[l-1])$ for $i = 1, \dots, N_p - 1$.
- 8: Compute $\{\tilde{w}_{t-1|T}^i\}_{i=1}^{N_p}$ using

$$\tilde{w}_{t-1|T}^i \propto w_{t-1}^i N(\tilde{x}_t | A[n-1]x_{t-1}^i, Q[n-1])$$

- 9: Draw $a_t^{N_p}$ with $P(a_t^{N_p} = j) \propto \tilde{w}_{t-1|T}^j$
 - 10: Set $x_t^{N_p} = \tilde{x}_t$
 - 11: Set $x_{1:t}^i = (x_{1:t-1}^{a_t^i}, x_t^i)$ for $i = 1, \dots, N_p$.
 - 12: Compute $w_t^i = N(y_t | W[n-1]\phi(x_t^i), \Sigma[n-1])$ for $i = 1, \dots, N_p$.
 - 13: **end for**
 - 14: Sample k with $P(k = i) \propto w_T^i$ and set $x_{1:T}[n] = x_{1:T}^k$
-

Appendix C

Hilbert Space Methods for Reduced-Rank Gaussian Processes

C.1 Kernels as Integral Operators

In this chapter¹, let operator be a continuous² and linear linear map between normed spaces(over the same field). An *integral operator*(or *transform*) is a mapping on functions

$$Tf(y) = \int_A k(x, y) f(x) d\mu(x)$$

where k is called kernel of T ³, and μ is a measure. A real symmetric kernel is such that $k(x, y) = k(y, x)$.

The *Gram* matrix K is the matrix whose ij entry is $K_{ij} = k(x_i, x_j)$, where $\mathcal{D} = \{x_i : i = 1, \dots, n\}$ is an input dataset. If k is a covariance function, then K is its covariance matrix.

A kernel is called positive semidefinite (PSD) when

$$\int k(x, y)f(x)f(y) d\mu(x) d\mu(y) \geq 0$$

. A PSD kernel gives rise to a PSD Gram matrix, for any n and \mathcal{D} and vice-versa.

¹We will follow closely the presentation in Rasmussen and Williams (2006)

²Some authors prefer the term *bounded*

³A different concept from that which is understood as $\ker T$

C.1.1 Stationary Kernels

For a zero-mean complex-valued process f , we define its covariance function as $k(x, y) = E(f(x)f(y)^*)$ (* corresponds to complex conjugation)

A *stationary* covariance function is a function of $\tau = x - y$, and it is usually written as $k(\tau)$, i.e. with a single argument.

By Bochner's Theorem, we can represent a stationary process as the Fourier transform of a positive finite measure.

Theorem C.1.1.1 *A complex-valued function k on \mathbb{R}^D is the covariance function of a weakly stationary mean square continuous complex-valued random process on \mathbb{R}^D if and only if it can be represented as*

$$k(\tau) = \int_{\mathbb{R}^D} e^{2\pi i \cdot s \cdot \tau} d\mu(s)$$

where μ is a positive finite measure.

If μ has a density $S(s)$, then S is called as the *spectral density* or *power spectrum* corresponding to k .

When this density exists, by the Wiener-Khinchin theorem, we know that

Theorem C.1.1.2 *If k and S satisfy the necessary conditions for Fourier inversion to be valid, then we have*

$$\begin{aligned} k(\tau) &= \int S(s) e^{2\pi i \cdot s \cdot \tau} ds \\ S(s) &= \int k(\tau) e^{-2\pi i \cdot s \cdot \tau} d\tau \end{aligned}$$

Since $k(0) = \int S(s) ds$ (process variance), the density of the process must also be integrable.

C.1.2 Eigenfunction Analysis of Kernels

In appendix A, we delved a bit on the Weight-Space view of GP, and on the notion of GP regression as a Bayesian linear regression with a number of basis functions. A possible choice of basis functions set is the set of eigenfunctions of the covariance function.

An eigenfunction ϕ of a kernel k with eigenvalue λ with respect to measure μ satisfies the integral equation

$$\int k(x, y) \phi(x) d\mu(x) = \lambda \phi(y).$$

Usually we can have an infinite number of eigenfunctions ϕ_i , and choose them such that they are orthonormalized, i.e. $\int \phi_i(x)\phi_j(x) d\mu(x) = \delta_{ij}$ (Kronecker delta).

The following theorem, called Mercer's theorem (a generalization of the original Mercer's theorem), tell us that we can write k as a function of eigenvalues and eigenfunctions.

Theorem C.1.2.1 *Let us be in a finite measure space, k be an a.e. bounded measurable function, such that its associated integral operator T_k is positive definite. Let the ϕ_i 's be square-integrable orthonormalized eigenfunctions of T associated to eigenvalues $\lambda_i > 0$. Then, we can state that*

- *the eigenvalues are absolutely summable*
- *$k(x, y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i(y)^*$ holds $\mu \times \mu$ a.e., and converging absolutely and uniformly in $\mu \times \mu$ a.e.*

Another useful theorem, known as the Karhunen–Loève theorem, giving us an expansion with the same name, and which uses the Mercer's theorem is the following:

Theorem C.1.2.2 *Let X_t be a zero-mean square-integrable stochastic process defined over a probability space, and indexed over a closed and bounded interval $[a, b]$, with continuous covariance function $k(s, t)$.*

Then $k(s, t)$ is a Mercer kernel⁴ and letting $\{\phi_j\}$ be an orthonormal basis on $L_2([a, b])$ formed by the eigenfunctions of T_k with respective eigenvalues λ_j , X_t admits the following representation

$$X_t = \sum_{j=1}^{\infty} Z_j \phi_j(t)$$

where the convergence is in L_2 , uniform in t and $Z_j = \int_a^b X_t \phi_j(t) dt$ and the random variables Z_j are zero mean, uncorrelated and have variance λ_j .

⁴A continuous symmetric positive definite kernel.

Appendix D

Derivations for Langevin type of Metropolis-Hastings Step

In this Appendix, we are interested in finding the derivative with respect to the new parameters $\tilde{\theta}$. This is achieved by composition, while differentiating a scalar function with respect to matrices.

One can observe from the marginal likelihood that

$$p(y_{1:T}, x_{1:T} | \tilde{\theta}) = p(x_1 | \tilde{\theta}) N(x_{2:T} | (I_{T-1} \otimes A_{\tilde{\theta}}) x_{1:T-1}, (I_{T-1} \otimes Q_{\tilde{\theta}})) \cdot N(y_{1:T} | \tilde{m}(x_{1:T}), \tilde{K}_{1:T})$$

Defining $\log(p(y_{1:T}, x_{1:T} | \tilde{\theta})) = l_{x_1} + l_x + l_y$, with

$$l_x = -\frac{1}{2} \log(\det(I_{T-1} \otimes Q_{\tilde{\theta}})) - \frac{1}{2} (x_{1:T} - (I_{T-1} \otimes A_{\tilde{\theta}}) x_{1:T-1})^\top (I_{T-1} \otimes Q_{\tilde{\theta}})^{-1} (x_{1:T} - (I_{T-1} \otimes A_{\tilde{\theta}}) x_{1:T-1})$$

and

$$l_y = -\frac{1}{2} \log(\det(K_M(x_{1:T}, x_{1:T}) \otimes M_1 + I_T \otimes \Sigma_{\tilde{\theta}})) - \frac{1}{2} (y_{1:T} - (I_T \otimes C_{\tilde{\theta}}) x_{1:T})^\top (K_M(x_{1:T}, x_{1:T}) \otimes M_1 + I_T \otimes \Sigma_{\tilde{\theta}})^{-1} (y_{1:T} - (I_T \otimes C_{\tilde{\theta}}) x_{1:T} - C_{*1:T}(\tilde{\theta}))$$

To reach the derivations of a scalar function with respect to a matrix¹, we will use the fact that $df(X) = \sum_k \sum_l \frac{\partial}{\partial X_{kl}} f(X) dX_{kl} = \text{Tr}((\frac{\partial}{\partial X} f(X))^\top dX)$, and some known facts which can be found on **PP12**, and **Gen17**

¹In all the derivations in this section, we will use the column notation, i.e., with v a vector and f a scalar function, $\frac{\partial}{\partial v} f$ is a column vector

To find $\frac{\partial}{\partial A} l_x$, let us define $v_A = x_{2:T} - (I_{T-1} \otimes A)x_{1:T-1}$, and notice that the term we are interested in can be rewritten in the form $-\frac{1}{2}v_A^\top (I_{T-1} \otimes Q)^{-1}v_A$. We know that $\partial f = \left(\frac{\partial}{\partial v_A} f\right)^\top \partial v_A$ and so $\left(\frac{\partial}{\partial v_A} f\right)^\top = -\frac{1}{2}2v_A^\top (I_{T-1} \otimes Q)^{-1}$. Hence, the differential w.r.t A only is

$$\begin{aligned}\partial(l_x) &= -(x_{2:T} - (I_{T-1} \otimes A)x_{1:T-1})^\top (I_{T-1} \otimes Q)^{-1} d(-(I_{T-1} \otimes A)x_{1:T-1}) \\ &= v_A^\top (I_{T-1} \otimes Q)^{-1} (vec(\partial A \tilde{X}_{1:T-1} I_{T-1}))\end{aligned}$$

where $\tilde{X}_{1:T-1}$ is defined such that it's conformable above and $vec(\tilde{X}_{1:T-1}) = x_{1:T-1}$. Let's also define W_{AQ} in a similar way, such that $vec(W_{AQ})^\top = v_A^\top (I_{T-1} \otimes Q)^{-1}$, with $\dim(W_{AQ}) = 2 \times (T-1)$ and so we obtain

$$\begin{aligned}\partial(l_x) &= Tr(W_{AQ}^\top \partial A \tilde{X}_{1:T-1} I_{T-1}) \\ &= Tr(\tilde{X}_{1:T-1} \Gamma_{AQ}^\top \partial A)\end{aligned}$$

and so,

$$\frac{\partial}{\partial A} l_x = W_{AQ} \tilde{X}_{1:T-1}^\top$$

Now, we must observe that the priors we will use were put on the economic parameters, not on A . Therefore, we must consider A as a function of the reparametrized economic parameters $\tilde{\theta}_{\text{econ}} = (\tilde{\beta}, \tilde{\delta}, \tilde{\phi}, \tilde{\rho}, \tilde{\alpha}, \tilde{\gamma})^\top$. So, abusing slightly our notation, let us write $A(\tilde{\theta}_{\text{econ}})$. Now notice that $\partial A(\tilde{\theta}_{\text{econ}}) = \left[\left(\frac{\partial a_{ij}}{\partial \tilde{\theta}_{\text{econ}}}(\tilde{\theta}_{\text{econ}})\right)^\top \partial \tilde{\theta}_{\text{econ}}\right] = \left[\left(\frac{\partial a_{ij}}{\partial \tilde{\theta}_{\text{econ}}}(\tilde{\theta}_{\text{econ}})\right)^\top\right] (I_2 \otimes \partial \tilde{\theta}_{\text{econ}})$.

So, we now know that $\partial l_x = Tr\left(\left(\frac{\partial}{\partial A} l_x\right)^\top \partial A\right) = Tr\left(H_A^\top (I_2 \otimes \partial \tilde{\theta}_{\text{econ}})\right)$, where $H_A^\top = \left(\frac{\partial}{\partial A} l_x\right)^\top \left[\left(\frac{\partial a_{ij}}{\partial \tilde{\theta}_{\text{econ}}}(\tilde{\theta}_{\text{econ}})\right)^\top\right]$. Now, $\partial l_x = Tr\left(H_A^\top (I_2 \otimes d\tilde{\theta}_{\text{econ}})\right) = Tr\left(vec(H_A)^\top vec(I_2 \otimes d\tilde{\theta}_{\text{econ}})\right) = Tr\left(vec(H_A)^\top (vec(I_2) \otimes I_6) d\tilde{\theta}_{\text{econ}}\right)$, which implies that

$$\frac{\partial l_x}{\partial A} \frac{\partial A}{\partial \tilde{\theta}_{\text{econ}}} = (vec(I_2) \otimes I_6)^\top vec(H_A).$$

To find $\frac{\partial l_x}{\partial Q} \frac{\partial Q}{\partial (q_i^2)}$, we notice that the differential w.r.t Q (only) is the sum of the differential of the two terms, so

$$\begin{aligned}\frac{\partial}{\partial Q} \left(-\frac{1}{2} \log(\det(I_{T-1} \otimes Q))\right) &= -\frac{1}{2} \frac{\partial}{\partial Q} \log((\det(I))^{\dim Q} (\det(Q))^{\dim I_{T-1}}) \\ &= -\frac{1}{2} (T-1) \frac{\partial}{\partial Q} \log(\det(Q)) = -\frac{1}{2} (T-1) (Q^{-1})^\top\end{aligned}$$

To simplify the notation, using v_A defined as above, the remaining term in l_x to derive by Q is $-\frac{1}{2}v_A^\top(I_{T-1} \otimes Q)^{-1}v_A$

$$\begin{aligned}\partial\left(-\frac{1}{2}v_A^\top(I_{T-1} \otimes Q)^{-1}v_A\right) &= -\frac{1}{2}v_A^\top\partial((I_{T-1} \otimes Q)^{-1})v_A \\ &= -\frac{1}{2}v_A^\top(-(I_{T-1} \otimes Q)^{-1})(I_{T-1} \otimes \partial Q)(I_{T-1} \otimes Q)^{-1}v_A \\ &= \frac{1}{2}v_A^\top(I_{T-1} \otimes Q)^{-1}\text{vec}(\partial Q W_Q I_{T-1})\end{aligned}$$

where W_{AQ} is as defined above, i.e. conformable with $\dim(W_{AQ}) = 2 \times (T-1)$ and with $\text{vec}(W_{AQ}) = (I_{T-1} \otimes Q)^{-1}v_A$.

Hence,

$$\begin{aligned}\partial(-v_A^\top(I_{T-1} \otimes Q)^{-1}v_A) &= \text{vec}(W_{AQ})^\top \text{vec}(\partial Q W_{AQ}) \\ &= \text{Tr}(W_{AQ}^\top \partial Q W_{AQ}) \\ &= \text{Tr}(W_{AQ} W_{AQ}^\top \partial Q)\end{aligned}$$

and so we obtain

$$\frac{\partial}{\partial Q}(-v_A^\top(I_{T-1} \otimes Q)^{-1}v_A) = \frac{1}{2}W_{AQ}W_{AQ}^\top$$

and thus we have:

$$\frac{\partial}{\partial Q}l_x = -\frac{1}{2}(T-1)(Q^{-1})^\top + \frac{1}{2}W_{AQ}W_{AQ}^\top$$

We must also observe that $\partial Q = \text{Diag}(\partial q_i^2) = \text{Diag}(e^{\tilde{q}_i^2} \partial \tilde{q}_i^2)$, hence $\partial l_x = \text{Tr}\left(\left(\frac{\partial}{\partial Q}l_x\right)^\top \partial Q\right) = \text{Tr}\left(\left(\frac{\partial}{\partial Q}l_x\right)^\top \text{Diag}(\partial q_i^2)\right) = \text{Tr}\left(\text{diag}\left(\left(\frac{\partial}{\partial Q}l_x\right)^\top \text{Diag}(e^{\tilde{q}_i^2})\right) \left[\partial \tilde{q}_i^2\right]\right)$, where $\text{Diag}(a_i)$ is a diagonal matrix with entries a_i , and $\text{diag}(A)$ is the main diagonal of a matrix A . Thus,

$$\frac{\partial l_x}{\partial Q} \frac{\partial Q}{\partial(\tilde{q}_i)} = \text{diag}\left(\left(\frac{\partial}{\partial Q}l_x\right)^\top \text{Diag}(e^{\tilde{q}_i})\right)$$

Now, we must look at l_y . Defining $U_{M\Sigma} = K_M \otimes M_1 + I_T \otimes \Sigma$, we will

first compute $\frac{\partial}{\partial M_1} l_y$

$$\begin{aligned}
\partial\left(-\frac{1}{2}\log(\det(K_M \otimes M_1 + I_T \otimes \Sigma))\right) &= \text{Tr}\left(-\frac{1}{2}U_{M\Sigma}^{-1}\partial(U_{M\Sigma})\right) \\
&= \text{Tr}\left(-\frac{1}{2}U_{M\Sigma}^{-1}(K_M \otimes \partial M_1)\right) \\
&= -\frac{1}{2}\text{vec}(U_{M\Sigma}^{-1})^\top \text{vec}(K_M \otimes \partial M_1) \\
&= -\frac{1}{2}\text{vec}(U_{M\Sigma}^{-1})^\top H_M \text{vec}(\partial M_1) \\
&= -\frac{1}{2}\text{vec}(\Gamma_M)^\top \text{vec}(\partial M_1) \\
&= \text{Tr}\left(-\frac{1}{2}\Gamma_M^\top \partial M_1\right)
\end{aligned}$$

where

$$H_M = \begin{bmatrix} K_M e_1 \otimes (e_1^\top \otimes I_D) \\ \vdots \\ K_M e_1 \otimes (e_D^\top \otimes I_D) \\ K_M e_2 \otimes (e_1^\top \otimes I_D) \\ \vdots \\ K_M e_2 \otimes (e_D^\top \otimes I_D) \\ K_M e_3 \otimes (e_1^\top \otimes I_D) \\ \vdots \\ K_M e_T \otimes (e_D^\top \otimes I_D) \end{bmatrix} = \begin{bmatrix} I_D \otimes K_M e_1 \\ \vdots \\ I_D \otimes K_M e_T \end{bmatrix} \otimes I_D$$

and Γ_M is such that $\text{vec}(\Gamma_M)^\top = \text{vec}(U_{M\Sigma}^{-1})^\top H_M$ and $\dim(\Gamma_l) = D \times D$. Therefore,

$$\frac{\partial}{\partial M_1} \left(-\frac{1}{2} \log(\det(M_1 \otimes K_M + I_T \otimes \Sigma)) \right) = -\frac{1}{2} \Gamma_M$$

Similarly to what we have done with l_x , to simplify the notation when calculating the second term in $\frac{\partial}{\partial M_1} l_y$, we will define $v_C = y_{1:T} - (I_T \otimes C)x_{1:T} - C_{*1:T}$. So, for the second term we compute

$$\begin{aligned}
\partial\left(-\frac{1}{2}v_C^\top (M_1 \otimes K_M + I_T \otimes \Sigma)^{-1} v_C\right) &= -\frac{1}{2}v_C^\top (-U_{M\Sigma}^{-1}) \partial U_{M\Sigma} U_{M\Sigma}^{-1} v_C \\
&= \frac{1}{2}v_C^\top U_{M\Sigma}^{-1} (K_M \otimes \partial M_1) U_{M\Sigma}^{-1} v_C \\
&= \frac{1}{2}\text{vec}(W_{M\Sigma})^\top \text{vec}(\partial M_1 W_{M\Sigma} K_M)
\end{aligned}$$

where $W_{M\Sigma}$ is defined by being conformable, i.e. $\dim(W_{M\Sigma}) = D \times T$ and $\text{vec}(W_{M\Sigma}) = U_{M\Sigma}^{-1}v_C$. Therefore, by symmetry

$$\partial(-v_C^\top(M_1 \otimes K_M + I_T \otimes \Sigma)^{-1}v_C) = \frac{1}{2}\text{Tr}(W_{M\Sigma}K_MW_{M\Sigma}^\top\partial M_1)$$

and so we have,

$$\frac{\partial l_y}{\partial M_1} = -\frac{1}{2}\Gamma_M + \frac{1}{2}W_{M\Sigma}K_MW_{M\Sigma}^\top$$

Let us now look at the derivative of l_y with respect to l , the length-scale parameter of the covariance function. From above, we have

$$\begin{aligned}\partial(-\frac{1}{2}v_C^\top(K_M \otimes M_1 + I_T \otimes \Sigma)^{-1}v_C) &= \frac{1}{2}(\text{vec}(W_{M\Sigma}))^\top \text{vec}(M_1W_{M\Sigma}\partial K_M) \\ &= \frac{1}{2}\text{Tr}(W_{M\Sigma}^\top M_1W_{M\Sigma}\partial K_M)\end{aligned}$$

Also, notice that

$$\begin{aligned}\partial K_M &= \left[\frac{\partial K_{ij}}{\partial l} \exp(\tilde{l}) \partial \tilde{l} \right] \\ &= \left[\exp\left(\frac{-r_{ij}}{l}\right) \cdot \frac{r_{ij}}{l^2} \exp(\tilde{l}) \right] \partial \tilde{l} \\ &= \left[\exp\left(\frac{-r_{ij}}{l}\right) \cdot \frac{r_{ij}}{l} \right] \partial \tilde{l}\end{aligned}$$

And so we have,

$$\frac{\partial}{\partial \tilde{l}}(-v_C^\top U_{M\Sigma}^{-1}v_C) = \text{Tr} \left(\frac{1}{2}W_{M\Sigma}^\top M_1W_{M\Sigma} \left[\exp\left(\frac{-r_{ij}}{l}\right) \cdot \frac{r_{ij}}{l} \right] \right)$$

For the remaining term as a function of l , we have

$$\partial\left(-\frac{1}{2}\log(\det(K_M \otimes M_1 + I_T \otimes \Sigma))\right) = -\frac{1}{2}\text{vec}(U_{M\Sigma}^{-1})^\top \text{vec}(\partial K_M \otimes M_1)$$

$$\begin{aligned}
\text{vec}(\partial K_M \otimes M_1) &= \begin{bmatrix} \partial(K_M)_{11} \text{Col}(M_1, 1) \\ \vdots \\ \partial(K_M)_{D1} \text{Col}(M_1, 1) \\ \partial(K_M)_{11} \text{Col}(M_1, 2) \\ \vdots \\ \partial(K_M)_{DD} \text{Col}(K_M, T) \end{bmatrix} \\
&= \left(I_T \otimes \begin{bmatrix} \vdots \\ I_T \otimes \text{Col}(M_1, i) \\ \vdots \end{bmatrix} \right) \text{vec}(\partial K_M)
\end{aligned}$$

where the $\text{Col}(M_1, i) = M_1 \cdot e_i$ are the i -th column of M_1 .

$$\begin{aligned}
\partial\left(-\frac{1}{2} \log(\det(K_M \otimes M_1 + I_T \otimes \Sigma))\right) &= -\frac{1}{2} \text{vec}(\Gamma_l)^\top \text{vec}(\partial K_M) \\
&= \text{Tr} \left(-\frac{1}{2} \Gamma_l^\top \partial K_M \right)
\end{aligned}$$

$$\text{where } \Gamma_l \text{ is defined by having } \text{vec}(\Gamma_l)^\top = \text{vec}(U_{M\Sigma}^{-1})^\top \left(I_T \otimes \begin{bmatrix} \vdots \\ I_T \otimes \text{Col}(M_1, i) \\ \vdots \end{bmatrix} \right)$$

and $\dim(\Gamma_l) = T \times T$

So, we have

$$\frac{\partial}{\partial l} \left(-\frac{1}{2} \log(\det(U_{M\Sigma})) \right) = \text{Tr} \left(-\frac{1}{2} \Gamma_l^\top \left[\exp\left(\frac{-r_{ij}}{l}\right) \cdot \frac{r_{ij}}{l} \right] \right)$$

And thus

$$\frac{\partial}{\partial l} (l_y) = \text{Tr} \left(-\frac{1}{2} \Gamma_l^\top \left[\exp\left(\frac{-r_{ij}}{l}\right) \cdot \frac{r_{ij}}{l} \right] \right) + \text{Tr} \left(\frac{1}{2} W_{M\Sigma} M_1 W_{M\Sigma}^\top \left[\exp\left(\frac{-r_{ij}}{l}\right) \cdot \frac{r_{ij}}{l} \right] \right)$$

For the $\frac{\partial l_y}{\partial \Sigma} \frac{\partial \Sigma}{\partial (\tilde{\sigma}_i^2)}$, we know, using the symmetry of $U_{M\Sigma}$,

$$\begin{aligned}
\partial \left(-\frac{1}{2} \log(\det(K_M \otimes M_1 + I_T \otimes \Sigma)) \right) &= \text{Tr} \left(-\frac{1}{2} U_{M\Sigma}^{-1} \partial U_{M\Sigma} \right) \\
&= -\frac{1}{2} \text{vec}(U_{M\Sigma}^{-1})^\top \text{vec}(I_T \otimes \partial \Sigma)
\end{aligned}$$

and noticing that

$$\text{vec}(I_T \otimes \partial \Sigma) = \left(\begin{bmatrix} I_D \otimes e_1 \\ \vdots \\ I_D \otimes e_T \end{bmatrix} \otimes I_D \right) \text{vec}(\partial \Sigma)$$

Therefore we have

$$\begin{aligned} \partial \left(-\frac{1}{2} \log(\det(K_M \otimes M_1 + I_T \otimes \Sigma)) \right) &= \text{Tr} \left(-\frac{1}{2} U_{M\Sigma}^{-1} \partial U_{M\Sigma} \right) \\ &= -\frac{1}{2} \text{vec}(U_{M\Sigma}^{-1})^\top \left(\begin{bmatrix} I_D \otimes e_1 \\ \vdots \\ I_D \otimes e_T \end{bmatrix} \otimes I_D \right) \text{vec}(\partial \Sigma) \\ &= \text{Tr} \left(-\frac{1}{2} (\Gamma_\Sigma)^\top \partial \Sigma \right) \end{aligned}$$

$$\text{Such that } \Gamma_\Sigma \text{ is defined by } \text{vec}(\Gamma_\Sigma)^\top = \text{vec}(U_{M\Sigma}^{-1})^\top \left(\begin{bmatrix} I_D \otimes e_1 \\ \vdots \\ I_D \otimes e_T \end{bmatrix} \otimes I_D \right),$$

and $\dim(\Gamma_\Sigma) = D \times D$.

Hence

$$\frac{\partial}{\partial \Sigma} \left(-\frac{1}{2} \log(\det(K_M \otimes M_1 + I_T \otimes \Sigma)) \right) = -\frac{1}{2} \Gamma_\Sigma$$

For the 2nd term of $\frac{\partial}{\partial \Sigma} l_y$,

$$\begin{aligned} \partial \left(-\frac{1}{2} v_C^\top (K_M \otimes M_1 + I_T \otimes \Sigma)^{-1} v_C \right) &= -\frac{1}{2} v_C^\top (-U_{M\Sigma})^{-1} (\partial U_{M\Sigma}) U_{M\Sigma}^{-1} v_C \\ &= \frac{1}{2} v_C^\top U_{M\Sigma}^{-1} (I_T \otimes \partial \Sigma) U_{M\Sigma}^{-1} v_C \\ &= \frac{1}{2} v_\Sigma^\top U_{M\Sigma}^{-1} \text{vec}(\partial \Sigma W_{M\Sigma} I_T) \end{aligned}$$

where $W_{M\Sigma}$ is defined as above, with $\text{vec}(W_{M\Sigma}) = U_{M\Sigma}^{-1} v_C$ and $\dim(W_{M\Sigma}) = D \times T$. Therefore, similarly to previous computations, we have

$$\begin{aligned} \partial \left(-\frac{1}{2} v_C^\top (K_M \otimes M_1 + I_T \otimes \Sigma)^{-1} v_C \right) &= \text{Tr} \left(\frac{1}{2} W_{M\Sigma}^\top \partial \Sigma W_{M\Sigma} I_T \right) \\ &= \text{Tr} \left(\frac{1}{2} W_{M\Sigma} W_{M\Sigma}^\top \partial \Sigma \right) \end{aligned}$$

And we have $\frac{\partial}{\partial \Sigma}(-v_{CR}^\top(K_M \otimes M_1 + I_T \otimes \Sigma)^{-1}v_{CR}) = W_{M\Sigma}W_{M\Sigma}^\top$, which implies

$$\frac{\partial l_y}{\partial \Sigma} = -\frac{1}{2}\Gamma_\Sigma + \frac{1}{2}W_{M\Sigma}W_{M\Sigma}^\top$$

and

$$\frac{\partial l_y}{\partial \Sigma} \frac{\partial \Sigma}{\partial(\tilde{\sigma}^2_i)} = \text{diag} \left(\left(\frac{\partial}{\partial \Sigma} l_y \right)^\top \text{Diag}(e^{\tilde{\sigma}^2_i}) \right).$$

To find $\frac{\partial}{\partial C} l_y$, we proceed in a very similar way to what we have done for $\frac{\partial}{\partial A} l_x$. The term we are interested is $-\frac{1}{2}v_C^\top U_{M\Sigma}^{-1}v_C$. We know that $df = \left(\frac{\partial}{\partial v_C} f\right)^\top dv_C$ and $\left(\frac{\partial}{\partial v_C} f\right)^\top = -\frac{2}{2}v_C^\top U_{M\Sigma}^{-1}$.

$$\begin{aligned} \partial(l_y) &= -v_C^\top U_{M\Sigma}^{-1} \partial(-(I_T \otimes C)x_{1:T}) \\ &= v_C^\top U_{M\Sigma}^{-1} (\text{vec}(\partial C \tilde{X}_{1T} I_T)) \end{aligned}$$

where \tilde{X}_{1T} is defined such that it's conformable above, with $\dim(\tilde{X}_{1T}) = D_x \times T$ since $\dim(C) = D \times D_x$, and $\text{vec}(\tilde{X}_{1T}) = x_{1:T}$. Using the definition of $W_{M\Sigma}$, so that $\text{vec}(W_{M\Sigma})^\top = v_C^\top U_{M\Sigma}^{-1}$, and so we obtain $\partial(l_y) = \text{Tr}(W_{M\Sigma}^\top \partial C \tilde{X}_{1T} I_T) = \text{Tr}(X_{1T} W_{M\Sigma}^\top \partial C)$.

And so

$$\frac{\partial}{\partial C} l_x = W_{M\Sigma} \tilde{X}_{1T}^\top$$

Since the prior influencing C was put on the economic parameters, we must proceed similarly to how we did for A . We know $\partial C_{4 \times 2}(\tilde{\theta}_{\text{econ}}) = \left[\left(\frac{\partial c_{ij}}{\partial \tilde{\theta}_{\text{econ}}} \right)^\top \right] (I_2 \otimes d\tilde{\theta}_{\text{econ}})$. So, we now know that $\partial l_y = \text{Tr} \left(\left(\frac{\partial}{\partial C} l_y \right)^\top dC \right) = \text{Tr} \left(H_C^\top (I_2 \otimes \partial \tilde{\theta}_{\text{econ}}) \right)$, where $H_C^\top = \left(\frac{\partial}{\partial C} l_y \right)^\top \left[\left(\frac{\partial c_{ij}}{\partial \tilde{\theta}_{\text{econ}}} (\tilde{\theta}_{\text{econ}}) \right)^\top \right]$.
Now,

$$\begin{aligned} \partial l_y &= \text{Tr} \left(H_C^\top (I_2 \otimes \partial \tilde{\theta}_{\text{econ}}) \right) = \text{Tr} \left(\text{vec}(H_C)^\top \text{vec}(I_2 \otimes \partial \tilde{\theta}_{\text{econ}}) \right) \\ &= \text{Tr} \left(\text{vec}(H_C)^\top (\text{vec}(I_2) \otimes I_6) \partial \tilde{\theta}_{\text{econ}} \right) \end{aligned}$$

and so we have that

$$\frac{\partial l_y}{\partial C} \frac{\partial C}{\partial \tilde{\theta}_{\text{econ}}} = (\text{vec}(I_2) \otimes I_6)^\top \text{vec}(H_C)$$

For finding the contribution of $\tilde{\theta}$ with respect to $\frac{\partial}{\partial C_*} l_y$ with $\dim(C_*) = D \times 1$, we first observe that

$$\partial l_y = \left(\frac{\partial l_y}{\partial v_C} \right)^\top \begin{bmatrix} -I_3 \\ \vdots \\ -I_3 \end{bmatrix} \partial C_* = -v_C^\top U_{M\Sigma}^{-1} \begin{bmatrix} -I_3 \\ \vdots \\ -I_3 \end{bmatrix} \left[\left(\frac{\partial}{\partial \tilde{\theta}_{\text{econ}}} (C_*)_i \right)^\top \right] \partial \tilde{\theta}_{\text{econ}}$$

so we have

$$\frac{\partial}{\partial C_*} l_y \frac{\partial}{\partial \tilde{\theta}} C_* = \left(v_C^\top U_{M\Sigma}^{-1} \begin{bmatrix} I_3 \\ \vdots \\ I_3 \end{bmatrix} \left[\left(\frac{\partial}{\partial \tilde{\theta}_{\text{econ}}} (C_*)_i \right)^\top \right] \right)^\top$$

Appendix E

Numerical Considerations

E.1 Alternative to Inverting a Matrix

Inverting a matrix may be a dangerous task, specially if the matrix is badly-conditioned, i.e. we are working at the edge of machine precision, which usually tends to appear more as the dimensions of the matrix increases.

Hence, it is praxis to use linear solver formulas to calculate expressions such as $CA^{-1}B$ or $CA^{-1}b$, since

$$Ax = b \Leftrightarrow x = A^{-1}b$$

In our context, since we are dealing, at least theoretically, with symmetric and Positive Definite matrices, one could use the Cholesky Decomposition/Method in the Mathematica command `LinearSolve`.

E.2 Guaranteeing Positive Definiteness and Symmetry

By working close to machine-precision, in certain steps, despite theory ensuring us that we were dealing with Symmetric Positive Definite (SPD) matrices, we needed to guarantee in a numerical way the matrix was SPD.

The rationale used in our work was, after the values were drawn, to find the closest (i.e. minimum), in the Frobenius norm sense, Positive Definite matrix. However, since that set is not closed, i.e. there is no minimum, just an infimum, we must be satisfied by finding a matrix indistinguishable from that up to machine precision.

From matrix X , we proceed as:

- 1 Find $Y = \frac{1}{2}(X + X^T)$, the closest symmetric matrix to X .

- 2 Take an eigendecomposition $Y = QDQ^\top$, and form the diagonal matrix $D_+ = \max(D, 0)$ (elementwise maximum).¹
- 3 Find the smallest $\epsilon > 0$ such that the CPU recognizes the following as a PD matrix: $Z = Q(D_+ + \epsilon I)Q^\top$. Theoretically, any positive value should be valid. However, due to machine precision, there is a lower bound.

Then just use matrix Z as your closest PD matrix.

If somehow, again by machine precision considerations, some matrix is not symmetric, we simply do $X := 0.5(X^\top + X)$.

¹Since M_1 is a small matrix, the potentially burdensome task of finding the eigendecomposition is relatively light.

Bibliography

- Adelino, M., A. Schoar, and F. Severino (2016). “Loan Originations and Defaults in the Mortgage Crisis: The Role of the Middle Class”. In: *The Review of Financial Studies* 29, pp. 1635–1670.
- Alvarez, I., J. Niemi, and M. Simpson (2014). “Bayesian inference for a covariance matrix”. In: *Proceedings of 26th Annual Conference on Applied Statistics in Agriculture* 26, pp. 71–82.
- Alvarez, M. A., L. Rosasco, and N. D. Lawrence (2012). “Kernels for Vector-Valued Functions: A Review”. In: *Foundation and Trends in Machine Learning* 4, pp. 195–266.
- An, S. and F. Schorfheide (2007). “Bayesian Analysis of DSGE Models”. In: *Econometric Reviews* 26, pp. 113–172.
- Andrieu, Christophe, Arnaud Doucet, and Roman Holenstein (2010). “Particle Markov Chain Monte Carlo Methods”. In: *Journal of Royal Statistical Society B* 72, pp. 269–342.
- Angrist, J. and A. Krueger (2001). “Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments”. In: *Journal of Economic Perspectives* 15, pp. 69–85.
- Arellano, M. (2003). *Panel Data Econometrics*. Oxford University Press.
- Arendt, P. D., D. W. Apley, and Wei Chen (2012). “Quantification of Model Uncertainty: Calibration, Model Discrepancy, and Identifiability”. In: *Journal of Mechanical Design* 134.
- Arendt, P. D., D. W. Apley, Wei Chen, et al. (2012). “Improving Identifiability in Model Calibration Using Multiple Responses”. In: *Journal of Mechanical Design* 134.
- Bayarri, M. J. et al. (2007). “A Framework for Validation of Computer Models”. In: *Technometrics* 49, pp. 138–154.
- Bernanke, B., M. Gertler, and S. Gilchrist (1999). “The financial accelerator in a quantitative business cycle framework”. In: *Handbook of Macroeconomics*. Ed. by K. Arrow and M. Intriligator. Vol. 1. Cambridge University Press. Chap. 12, pp. 1341–1393.

- Blanchard, O. (2016). *Do DSGE models have a future? - Peterson Institute for International Economics*. URL: <https://www.piie.com/publications/policy-briefs/do-dsge-models-have-future>.
- (2018). “On the future of macroeconomic models”. In: *Oxford Review of Economic Policy* 34, pp. 43–54.
- Boivin, J. and M. Giannoni (2006). “DSGE Models in a Data-Rich Environment”. In: *NBER Working Papers* 12772.
- Brynjarsdóttir, J. and A. O’Hagan (2014). “Learning about Physical Parameters: The Importance of Model Discrepancy”. In: *Inverse Problems* 30.
- Cai, M. et al. (2019). “Online Estimation of DSGE Models”. In: *PIER Working Paper No. 19-014* 28.
- Canova, F. (2014). “Bridging Cyclical DSGE Models and the Raw Data”. In: *Journal of Monetary Economics* 67, pp. 1–15.
- Canova, F. and C. Matthes (2017). “Composite Likelihood Approach for Dynamic Structural Models”. In: *Federal Reserve Bank Richmond - Working Paper*.
- Cappé, Olivier, Simon J. Godsill, and Eric Moulines (2007). “An Overview of Existing Methods and Recent Advances in Sequential Monte Carlo”. In: *Proceedings of the IEEE* 95, pp. 899–924.
- Chari, V. V., P. J. Kehoe, and E. McGrattan (2007). “Business Cycle Accounting”. In: *Econometrica* 75, pp. 781–836.
- Christiano, L. J., M. S. Eichenbaum, and M. Trabandt (2018). “On DSGE Models”. In: *Journal of Economic Perspectives* 32, pp. 113–140.
- Clarke, B. (2003). “Comparing bayes model averaging and stacking when model approximation error cannot be ignored”. In: *The Journal of Machine Learning Research* 4, pp. 683–712.
- Consolo, A., C. Favero, and A. Paccagnini (2009). “On the Statistical Identification of DSGE Models”. In: *Journal of Econometrics* 150, pp. 99–115.
- Corradi, V. and N. R. Swanson (2007). “Evaluation of Dynamic Stochastic General Equilibrium Models Based on Distributional Comparison of Simulated and Historical Data”. In: *Journal of Econometrics* 136, pp. 699–723.
- Dashti, M. and A. M. Stuart (2015). *The Bayesian Approach to Inverse Problems*. URL: <https://arxiv.org/abs/1302.6989>.
- DeJong, D. and C. Dave (2011). *Structural Macroeconometrics. Second Edition*. Princeton University Press.
- Diamond, P. (1965). “National debt in a neoclassical growth model”. In: *American Economic Review* 55, pp. 1126–1150.
- Doucet, Arnaud and Nando de Freitas (2001). “An Introduction to Sequential Monte Carlo Methods”. In: *Sequential Monte Carlo Methods in Practice*.

- Ed. by R. Ghanem, D. Higdon, and H. Owhadi. Springer. Chap. 1, pp. 3–14.
- Doucet, Arnaud and Adam M. Johansen (2008). *A Tutorial on Particle Filtering and Smoothing: Fifteen years later*. URL: https://www.stats.ox.ac.uk/~doucet/doucet_johansen_tutorialPF2011.pdf.
- Fragoso, T., W. Bertoli, and F. Louzada (2018). “Bayesian Model Averaging: A Systematic Review and Conceptual Classification”. In: *International Statistical Review* 86, pp. 1–28.
- Frigola-Alcalde, R. (2015). *Bayesian Time Series Learning with Gaussian Processes - PhD Thesis*. University of Cambridge.
- Frigola-Alcalde, R. et al. (2013). “Bayesian Inference and Learning in Gaussian Process State-Space Models with Particle MCMC”. In: *Advances in Neural Information Processing Systems* 26, pp. 3156–3164.
- Gentle, J. E. (2017). *Matrix Algebra: Theory, Computations and Applications in Statistics*. Springer.
- Guerron-Quintana, P. A. (2010). “What you Match Does Matter: The Effects of Data on DSGE Estimation”. In: *Journal of Applied Econometrics* 25, pp. 774–804.
- Gupta, A. K. and D. K. Nagar (1999). *Matrix Variate Distributions*. CRC Press.
- Gürkaynak, R. S. and C. Tille, eds. (2017). *DSGE Models in the Conduct of Policy: Use as intended*. VoxEU.org e-Book. URL: <https://voxeu.org/content/dsge-models-conduct-policy-use-intended>.
- Haan, Wouter J. Den and Thomas Drechsel (2018). “Agnostic Structural Disturbances: Detecting and Reducing Misspecification in Empirical Macroeconomic Models”. In: *CEPR Discussion Paper* 13145.
- Hansen, G. D. (1985). “Indivisible labor and the business cycle”. In: *Journal of Monetary Economics* 16, pp. 309–327.
- Harenberg, Daniel et al. (2019). “Uncertainty Quantification and Global Sensitivity Analysis for Economic Models”. In: *Quantitative Economics* 10, pp. 1–41.
- Ingram, B. and C. Whiteman (1994). “Supplanting the Minnesota Prior - Forecasting Macroeconomics Time Series using Real Business Cycle Model Priors”. In: *Journal of Monetary Economics* 34, pp. 497–510.
- Ingram, B. F., N. R. Kocherlakota, and N. E. Savin (1994). “Explaining Business Cycles: A Multiple-Shock Approach”. In: *Journal of Monetary Economics* 34, pp. 415–428.
- Inoue, A., C.-H. Kuo, and B. Rossi (2019). “Identifying the Sources of Model Misspecification”. In: *Journal of Monetary Economics*.
- International Monetary Fund (2009). “World Economic Outlook: Crisis and Recovery”. In: URL: <https://www.imf.org/en/Publications/WEO/>

[Issues/2016/12/31/World-Economic-Outlook-April-2009-Crisis-and-Recovery-22575](#).

- Ireland, Peter (2004). “A Method for Taking Models to the Data”. In: *Journal of Economic Dynamics and Control* 4, pp. 195–266.
- Jiang, Z. et al. (2017). “Multi-Response Approach to Improving Identifiability in Model Calibration”. In: *Handbook of Uncertainty Quantification*. Ed. by R. Ghanem, D. Higdon, and H. Owhadi. Springer. Chap. 4, pp. 69–129.
- Joseph, V. R. and H. Yan (2015). “Engineering-Drive Statistical Adjustment and Calibration”. In: *Technometrics* 57, pp. 257–267.
- Kennedy, M. C. and A. O’Hagan (2001). “Bayesian Calibration of Computer Models”. In: *Journal of the Royal Statistical Society B* 63, pp. 425–464.
- Kocherlakota, N. R. (2007). “Model Fit and Model Selection”. In: *Federal Reserve Bank of St. Louis* (July), pp. 349–360.
- Lindsten, F., M. I. Jordan, and T. S. Schön (2014). “Particle Gibbs with Ancestor Sampling”. In: *Journal of Machine Learning Research* 15, pp. 2145–2184.
- Liu, F., M.J. Bayarri, and J.O. Berger (2009). “Modularization in Bayesian Analysis, with Emphasis on Analysis of Computer Models”. In: *Bayesian Analysis* 4, pp. 119–150.
- Liu, H., J. Cai, and Yew-Soon Ong (2018). “Remarks on Multi-Output Gaussian Process Regression”. In: *Knowledge-Based Systems* 144, pp. 102–121.
- Loeppky, J., D. Bingham, and W. Welch (2006). *Computer Model Calibration or Tuning in Practice. Technical Report*. University of British Columbia.
- Lucas, R. (1976). “Econometric policy evaluation: A critique”. In: *Carnegie-Rochester Conference Series on Public Policy* 1, pp. 19–46.
- (2003). “Macroeconomic Priorities”. In: *American Economic Review* 93, pp. 1005–1014.
- Monti, F. (2015). “Can a Data-Rich Environment Help Identify the Sources of Model Misspecification?” In: *Working Paper - Bank of England*.
- Morris, S. D. (2016). “VARMA Representation of DSGE Models”. In: *Economic Letters* 138, pp. 30–33.
- Muellbauer, J. (2018). “The future of macroeconomics”. In: *The future of central banking - Festschrift in honour of Vítor Constâncio*. European Central Bank. Chap. 2, pp. 6–45.
- Negro, M. Del and F. Schorfheide (2004). “Priors from General Equilibrium Models for VARs”. In: *International Economic Review* 45, pp. 643–673.
- (2009). “Monetary Policy Analysis with Potentially Misspecified Models”. In: *American Economic Review* 99, pp. 1415–1450.

- Negro, M. Del and F. Schorfheide (2013). “DSGE Model-Based Forecasting”. In: *Handbook of Economic Forecasting*. Ed. by G. Elliott and A. Timmermann. Vol. 2. Cambridge University Press. Chap. 2, pp. 57–140.
- Negro, M. Del, F. Schorfheide, et al. (2005). “On the Fit and Forecasting Performance of New-Keynesian Models. Working Paper Series 491”. In: *European Central Bank*.
- Paccagnini, A. (2017). “Dealing with Misspecification in DSGE Models: A Survey. MPRA Paper 82914”. In: *University Library of Munich*.
- Paulo, R., G. Garcia-Donato, and J. Palomo (2012). “Calibration of Computer Models with Multivariate Output”. In: *Computational Statistics and Data Analysis* 56, pp. 3959–3974.
- Plumlee, M. and V.R. Joseph (2018). “Orthogonal Gaussian Process Models”. In: *Statistica Sinica* 28, pp. 601–619.
- Poudyal, N. and A. Spanos (2016). “Model Validation and DSGE Modelling”. In: *Working Paper*.
- Quinn, M. (2003). *Parallel Programming in C with MPI and OpenMP*. McGraw-Hill Education.
- Rasmussen, C. and C. Williams (2006). *Gaussian Processes for Machine Learning*. MIT Press.
- Ratto, M. (2008). “Analysing DSGE Models with Global Sensitivity Analysis”. In: *Computational Economics* 31, pp. 115–139.
- Reich, B. J. and J. S. Hodges (2010). “Adding Spatially-Correlated Errors can mess up the Fixed Effect you love”. In: *Biometrics* 64, pp. 325–334.
- Reich, B. J., J. S. Hodges, and V. Zadnik (2006). “Effects of Residual Smoothing on the Posterior of the fixed effects in disease-mapping models”. In: *Biometrics* 62, pp. 1197–1206.
- Romer, P. (2016). *The Trouble with Macroeconomics - Commons Memorial Lecture of the Omicron Delta Epsilon Society. Forthcoming in The American Economist*. URL: <https://paulromer.net/trouble-with-macroeconomics-update/WP-Trouble.pdf>.
- Scheidegger, Simon and Ilias Bilonis (2019). “Machine Learning for High-Dimensional Dynamic Stochastic Economies”. In: *Journal of Computational Science* 33, pp. 68–82.
- Schmitt-Grohé, S. and M. Uribe (2012). “What’s News in Business Cycles”. In: *Econometrica* 80, pp. 2733–2764.
- Schorfheide, F. (2013). “Estimation and Evaluation of DSGE Models: Progress and Challenges”. In: *Advances in Economics and Econometrics: Tenth World Congress*. Ed. by D. Acemoglu, M. Arellano, and E. Dekel. Cambridge University Press. Chap. 5, pp. 184–230.

- Schorfheide, F., K. Sill, and M. Kryskho (2010). “DSGE Model-Based Forecasting of NonModelled Variables”. In: *International Journal of Forecasting* 26, pp. 348–373.
- Smets, F. and R. Wouters (2003). “An Estimated Dynamic Stochastic General Equilibrium Model of the Euro Area”. In: *Journal of the European Economic Association* 1, pp. 1123–1175.
- (2007). “Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach”. In: *American Economic Review* 97, pp. 586–606.
- Soize, C. (2018). *Uncertainty Quantification: An Accelerated Course with Advanced Applications in Computational Engineering*. Springer.
- Solin, A. and S. Särkkä (2014). *Hilbert Space Methods for Reduced-Rank Gaussian Process Regression*. URL: <https://arxiv.org/abs/1401.5508>.
- Steel, M. (2019). *Model Averaging and its Use in Economics*. URL: <https://arxiv.org/abs/1709.08221>.
- Stuart, A. M. (2010). “Inverse Problems: A Bayesian Perspective”. In: *Acta Numerica* 19, pp. 451–559.
- (2014). *Uncertainty Quantification in Bayesian Inversion - ICM Talk*. URL: <https://homepages.warwick.ac.uk/~masdr/TALKS/stuartICM.pdf>.
- Sullivan, T. J. (2015). *Introduction to Uncertainty Quantification*. Springer.
- Svensson, A. et al. (2016). “Computationally Efficient Bayesian Learning of Gaussian Process State-Space Models”. In: *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*.
- The Financial Crisis Inquiry Commission (2011). *The Financial Crisis Inquiry Report - Final Report of the National Commission on the Causes of the Financial and Economic Crisis in the United States*. Tech. rep. US Government Printing Office.
- Tovar, C. (2008). “DSGE Models and Central Banks”. In: *Bank of International Settlements - Working Paper*.
- Wieland, V. et al. (2012). “A new comparative approach to macroeconomic modeling and policy analysis”. In: *Journal of Economic Behavior and Organization* 83, pp. 523–541.
- Wills, A. et al. (2012). “Estimation of Linear Systems using a Gibbs Sampler”. In: *Proceedings of the 16th IFAC Symposium on System Identification*, pp. 203–208.